G. Klimov

Probability Theory and Mathematical Statistics

Probability Theory and Mathematical Statistics

Г. П. Климов

Теория вероятностей и математическая статистика

ИЗДАТЕЛЬСТВО МОСКОВСКОГО УНИВЕРСИТЕТА

G. Klimov

Probability Theory and Mathematical Statistics

Translated from the Russian
by
Oleg Efimov

MIR PUBLISHERS MOSCOW

First published 1986
Revised from the 1983 Russian edition

На английском языке

© Издательство Московского университета, 1983
© English translation, Mir Publishers, 1986

Contents

List of Abbreviations 9

Preface 10

Part 1. Probability 11

Introduction 11

Chapter 1. **Discrete Probability Space** 11
§ 1. Event and Probability 11
§ 2. Probability Space 13
§ 3. Independence of Events 14
§ 4. Conditional Probability 15
§ 5. Probability of Union of Events 16
§ 6. Composite Probability Formula 17
§ 7. Bayes Formula 17
§ 8. Examples of Distributions 18
§ 9. Random Variable. Mathematical Expectation 19
§ 10. Independence of Random Variables 21
§ 11. Additive and Multiplicative Properties of Mathematical Expectation 22
§ 12. Moments 23
§ 13. Variance Additivity for Independent RRV's 24
§ 14. Generating Function 24
§ 15. Binomial Distribution 25
§ 16. Poisson Distribution 26
§ 17. Negative Binomial Distribution 26
§ 18. Approximation of Binomial by Poisson Distribution 27
§ 19. Indicator of Event 30
§ 20. Markov Inequality 30
§ 21. Law of Large Numbers 31
§ 22. Central Limit Theorem 34

Chapter 2. **Probability Space** 37
§ 1. Definition of Probability Space 37
§ 2. Construction of Probability Space 38
§ 3. Complete Probability Space 41
§ 4. Continuous Probability Space 42
§ 5. Definition of Random Variable 45
§ 6. Operations over RV's 46
§ 7. Equivalence and Almost Certain Convergence of RV's 47
§ 8. Independence 48
§ 9. Mathematical Expectation, Distribution Function, Distribution Density 51
§ 10. Characteristic Functions 54
§ 11. One-Dimensional Normal Distribution 58

§ 12. Gamma Distribution 60
§ 13. Beta and Other Distributions 61
§ 14. Uniform Distribution 64
§ 15. VRV Characteristic Function 67
§ 16. Multidimensional Normal Distribution 68
§ 17. Markov Inequality 72
§ 18. Jensen Inequality 72
§ 19. Kolmogorov Inequality 73
§ 20. Moment Inequalities 74
§ 21. Conditional Expectation 76
§ 22. Conditional Probability 80

Supplement: Subjective Probability 86

Problems 91

Part 2. Stochastic Processes 105

Introduction: Definition of Random Process. Kolmogorov Theorem on Construction of Processes from Partial Distributions 105

Chapter 3. Discrete- and Continuous-Time Markov Chains 111
§ 1. Definition of Discrete- and Continuous-Time Markov Chains 111
§ 2. Examples of Homogeneous Markov Chains 114
§ 3. Limiting Distribution of Markov Chains 119
§ 4. Proofs 121
§ 5. Law of Large Numbers for Markov Chains 124
§ 6. Structure of Homogeneous Continuous-Time Markov Process with Finite State Space 127
§ 7. Structure of Homogeneous Continuous-Time Markov Process with Countable State Space 132

Chapter 4. Renewal Processes 140
§ 1. Definition of Renewal Process 140
§ 2. Elementary Renewal Theorem 142
§ 3. Blackwell Theorem 143
§ 4. Key Renewal Theorem 144

Chapter 5. Regenerative Processes 148
§ 1. Definition of Regenerative Process 148
§ 2. Limit Theorem for Regenerative Process 150
§ 3. Law of Large Numbers for Regenerative Processes 154

Chapter 6. Elements of Random Analysis 159
§ 1. Types of RV Convergence 159
§ 2. Equivalent Definitions of RV Convergence and Basic Properties 160
§ 3. Prokhorov Theorem and Its Applications 161
§ 4. Stochastic Continuity 165
§ 5. Stochastic Integrals of Non-Random Functions 166
§ 6. Orthogonal Measure and Integral with Values in Hilbert Space 171
§ 7. Differentiation and Integration of Mathematical Expectation of Processes 179

Chapter 7. Gaussian Processes 181
§ 1. Definition of Gaussian Process 181
§ 2. Wiener Process 181
§ 3. Non-Homogeneous Wiener Process 183

Chapter 8. Stationary Processes 186
§ 1. Definition of Stationary Process (in Wide Sense). Examples 186
§ 2. Representation of Stationary Process in Terms of Process with Uncorrelated Increments 188
§ 3. Differentiation of Stationary Processes 191
§ 4. Integration of Stationary Processes 195

Supplement: Regenerative Processes with Dependent Regeneration Cycles of Markov Type 202
§ 1. Semi-Markov Renewal Process 202
§ 2. Sufficient Condition for SMRP Moment Existence 203
§ 3. Elementary SMRP Renewal Theorem 204
§ 4. SMRP Blackwell Theorem 207
§ 5. SMRP Key Theorem of Renewal Theory 208
§ 6. Limit Theorem for Regenerative Processes with Dependent Regeneration Cycles of Markov Type 209
§ 7. Example of Application of Regenerative Processes with Dependent Regeneration Cycles of Markov Type 214

Problems 216

Part 3. Statistics 229

Introduction 229

§ 1. Posing Statistical Problem 229
§ 2. Risk Function 231
§ 3. Strategy Ordering 232

Chapter 9. Mathematical Statistics: Basic Concepts 234
§ 1. Statistical Structure 234
§ 2. Statistic 235
§ 3. Sufficient Statistic 235
§ 4. Complete Statistic 237
§ 5. Free Statistic 238
§ 6. Sufficient Condition for Existence of Decision Rule with Uniformly Minimum Risk 239

Chapter 10. Point Estimation: Cramér-Rao Theory 243
§ 1. MVUE 243
§ 2. MVUE Uniqueness 245
§ 3. Cramér-Rao Inequality 245
§ 4. Efficient Estimators 247
§ 5. Cramér-Rao Inequality Improvement. Bhattacharya Inequality 250
§ 6. Multidimensional Analogue of Cramér-Rao Inequality 252
§ 7. Multidimensional Analogue of Bhattacharya Inequality 254
§ 8. Expressing MVUE in Terms of Sufficient Statistic 257

Chapter 11. Point Estimation: Maximum Likelihood Method 257
§ 1. Maximum Likelihood Estimator 257
§ 2. Consistent Estimators 261
§ 3. MLE Consistency 261
§ 4. MLE Asymptotic Normality 265

Chapter 12. Point Estimation: Least Squares Method 270
§ 1. Examples 270
§ 2. Normal Equation 271

§ 3. Gauss-Markov Theorem 272
§ 4. Linear Regression Model 274
§ 5. Normal Error Distribution 277

Chapter 13. Interval Estimation 279
§ 1. Definition 279
§ 2. Interval Maximum Likelihood Estimator 281
§ 3. Likelihood Ratio Principle 282
§ 4. Likelihood Ratio Principle Asymptotic Property 282
§ 5. Generalized Likelihood Ratio Principle 283
§ 6. Technique of Likelihood Ratio Principle Application 287

Chapter 14. Linear Model Multidimensional Analysis 288
§ 1. Introduction 288
§ 2. Assumptions *re* Space of Parameters $\theta = (\mu, a)$ 289
§ 3. MLE Property 291
§ 4. Estimating Function 291
§ 5. Interval MLE for $\theta = (\mu, a)$ 292
§ 6. Independence of $\hat{\mu}$ and $\hat{a}$ 293
§ 7. Estimating μ and a Separately 297

Chapter 15. Statistical Hypothesis Testing 300
§ 1. Hypothesis and Test 300
§ 2. Power Function. Test Ordering 301
§ 3. Fundamental Neyman-Pearson Lemma 301

Chapter 16. Sequential Analysis 302
§ 1. Problem Posing 302
§ 2. Probability Ratio Sequential Test for s Hypotheses 304
§ 3. Average Sample Size Finiteness in Using PRST 308

Problems 310

Index of Definitions 330

List of Abbreviations

a.c.	almost certainly, almost certain
a.e.	almost everywhere
as.	asymptotically
CE	consistent estimator
CF	characteristic function
d.	distribution
DF	distribution function
DR	decision rule
EE	efficient estimator
i.i.d.	independent and identically distributed
LF	likelihood function
LLN	law of large numbers
LSM	least squares method
ME	mathematical expectation
MLE	maximum likelihood estimator
MVUE	minimum variance unbiased estimator
P.	probability
PRST	probability ratio sequential test
RRV	real random variable
RF	random function
RV	random variable
SDT	sequential decision test
s.m.	square mean
SMP	semi-Markov process
SMRP	semi-Markov renewal process
UE	unbiased estimator
u.m.p.	uniformly most powerful
VRV	vector-valued random variable
w.	weakly

Preface

The contents of the textbook is composed of the lectures in probability theory and mathematical statistics for students of the Faculty of Computational Mathematics and Cybernetics of Moscow University. Probability theory and mathematical statistics are read for three terms (fifth, sixth and seventh) once a week, accompanied by seminar classes (in the fifth term, it is one class a week, while in the sixth and seventh, one class every other week).

The lectures are, therefore, broken termwise into three parts, viz., **Probability**, **Stochastic Processes** and **Statistics.** Two hundred problems given at seminar classes are distributed so that Part 1 contains one hundred of them, with Parts 2 and 3 including fifty problems each (i.e., in proportion to the frequency of seminars).

In Part 1 **Probability**, the accent is on the following.

1. In each example and each problem, a probability space is supposed to be chosen first. In some cases, several probability spaces may be selected for the same problem. We emphasize that probability theory starts only with choosing a probability space, and it is the probability space that is the subject matter of probability theory.
2. At the initial stage of the study of probability theory, combinatorial problems are not cultivated, since the cult of such problems cuts the ground on which probability theory is constructed from under the student's feet. Besides, this creates an impression of certain primitiveness and artificiality.
3. The discrete probability space is investigated; we give many statements often formulated only for a general probability space, but, generally, with proofs not altering (e.g., Khinchine law of large numbers).

In Part 2 **Stochastic Processes**, the choice of material nontraditional. We discuss continuous-time Markov processes with countable state space, renewal and regenerative processes, for it is random processes that most students will subsequently apply. Note, in addition, that the statements and proofs of ergodic theorems for Markov chains do not depend on (either discrete or continuous) time. Little attention is paid to the central limit theorems of probability theory (simplest central limit theorem is given as an example only in connection with the Prokhorov theorem applications).

In Part 3 **Statistics**, the basic concepts of mathematical statistics are first defined (statistical structure, negligibility, sufficient statistic, complete statistic, free statistic). Then only the theory of point estimation is given in detail (efficient estimators, MLE's, LSM-estimators), and also that of interval estimation (of maximum likelihood on the basis of likelihood ratio principle). The theory of hypothesis testing is recounted quite briefly (students who specialize in mathematical statistics then make up for this gap). Several proofs and treatment of the material may seem unhabitual to the specialist; however, they are assimilated easier by the beginners.

Part 1

Probability

Introduction

Probability theory is mathematical analysis of the concept of random experiment. Event and probability are basic notions of the theory. In formalizing these concepts, we proceed from the set-theoretic model of the theory of probability, suggested by A. N. Kolmogorov in 1929.

It is assumed that the result, or outcome, of a random experiment cannot be determined. Let $\Omega = \{\omega\}$ be the set of all outcomes, and A a certain event related to the experiment under consideration. It is natural to assume that the outcome of an experiment may indicate whether or not the event A has occurred. Therefore, A can be identified with a certain collection of outcomes, i.e., assumed to be a subset of elements from Ω. The elements $\omega \in \Omega$ can then be called elementary events. In this case, the union, intersection and complement of events coincide with the corresponding set-theoretic operations, and the class of events which are of interest to us must be closed with respect to these operations, i.e., the union, intersection, and complement of events are also events.

But probability is a numerical characteristic (function) of a class of events, with properties that should be similar to those of the frequency with which an event occurs.

Before turning to the formalization and investigation of these concepts under general assumptions, in the first chapter, we shall consider a special, but rather frequent case where the set of elementary events is at most countable.

Chapter 1

Discrete Probability Space

§ 1. Event and Probability

Let $\Omega = \{\omega\}$ be a finite or countable set. We call the elements of the set Ω **elementary events**, and the set Ω the **space of elementary events.**

Any subset A of the set Ω will be called an **event**. Denote by S the class of subsets of the set Ω, i.e., S is the class of events.

Further, let each $\omega \in \Omega$ be associated with a number $P(\omega)$ such that

$$P(\omega) \geqslant 0, \quad \sum_{\omega \in \Omega} P(\omega) = 1.$$

The number $P(\omega)$ is called the **probability of an elementary event** ω. By the probability of an event A, the number

$$P(A) = \sum_{\omega \in A} P(\omega)$$

is understood.

Example 1. Let

$$\Omega = \{\omega_1, \ldots, \omega_N\}, \qquad P(\omega_k) = 1/N \ \forall k,$$

then

$$P(A) = \frac{r}{N} = \frac{|A|}{|\Omega|},$$

where $r = |A|$ is the number of elementary events contained in the event A. This is the classical model of probability theory.

Example 2. When tossing a fair coin twice, we can assume that the space of elementary events is

$$\Omega = \{HH, HT, TH, TT\},$$

where H means "head", and T "tail". In tossing a coin n times, the space Ω of elementary events can be assumed to consist of elementary events of the form

$$\omega = (\varepsilon_1, \ldots, \varepsilon_n),$$

where $\varepsilon_i = H$ or T for $\forall i$. In this case (see Example 1), $N = 2^n$. Without loss of generality, we can assume that $\varepsilon_i = 0$ or 1 for $\forall i$. Consider the event

$$A = \{(\varepsilon_1, \ldots, \varepsilon_n): \varepsilon_1 + \ldots + \varepsilon_n = k\}$$

for a certain $k = 0, 1, \ldots, n$. Then

$$|A| = {}^nC_k, \text{ and } P(A) = {}^nC_k/2^n.$$

Note that the event A is "occurrence of the tail k times when tossing a coin n times".

Example 3 (*showing that the space of elementary events cannot be always regarded as finite*). Suppose two Persons I and II toss a coin in turn, with Person I beginning, until the head is obtained. The winner is the one who makes the last move. The possible states of the game are

$$\omega_1 = H,$$
$$\omega_2 = TH,$$
$$\omega_3 = TTH,$$
$$\ldots\ldots\ldots$$

Let $\Omega = \{\omega_1, \omega_2, \ldots\}$, $P(\omega_k) = 2^{-k}$. It is clear that $P(\Omega) = 1$. $A = \{\omega_{2k+1}: k \geqslant 0\}$ is the event "win of Person I." Then the probability of the win of Person I is

$$P(A) = \sum_{k \geqslant 0} P(\omega_{2k+1}) = \sum_{k \geqslant 0} 2^{-(2k+1)} = 2/3.$$

§ 2. Probability Space

In the preceding section, the three objects were introduced, viz.,

Ω, a finite or a countable set (space of elementary events),

S, the class of all (!) subsets of the set Ω (class of events),

and

P, a numerical function defined on S (where $P(A)$ is the probability of an event $A \in S$).

Note that the following properties are fulfilled:

A1. The class S is a σ-algebra, i.e.,

(1) $A \in S$ implies $A^c = \Omega \setminus A \in S$,

(2) $A_i \in S$ for $i = 1, 2, \ldots$ implies $\cup A_i \in S$, $\cap A_i \in S$.

A2. The numerical function P defined on S is

(1) normalized, i.e., $P(\Omega) = 1$,

(2) positive semi-definite, i.e., $P(A) \geqslant 0$ for all $A \in S$, and

(3) σ-additive, i.e., $P(\sum A_i) = \sum P(A_i)$, where $\sum A_i = \cup A_i$ if $A_i \cap A_j = \varnothing$, with $i \neq j$. Two events A and B are said to be **mutually exclusive** if $A \cap B \neq \varnothing$.

The triplet of objects (Ω, S, P) is called a **probability space.** Note that the main object for study in probability theory is the probability space. We have defined this space for the case where Ω is at most a countable set.

To specify a probability space (in the discrete case), it suffices to specify a set $\Omega = \{\omega_k\}$ and a mapping $\omega_k \to p_k = P(\omega_k)$ such that $p_k \geqslant 0$ and $\sum p_k = 1$, which is often convenient to do in the form of a probability distribution table, e.g.,

Ω	ω_1	ω_2	$\ldots$	ω_N
P	p_1	p_2	$\ldots$	p_N

It follows (show!) from the given properties A1 and A2 that, in particular,

(1) $\varnothing \in S$, and $P(\varnothing) = 0$,

(2) for two events A and B, the inclusion $A \subset B$ implies $P(A) \leqslant P(B)$,

(3) $P(A \cup B) = P(A) + P(B) - P(AB)$,

(4) $P(\cap A_k) \leqslant \sum P(A_k)$,

(5) if $A_n \uparrow A$, then $P(A) = \lim_{n \to \infty} P(A_n)$, and

(6) if $A_n \downarrow A$, then $P(A) = \lim\limits_{n\to\infty} P(A_n)$,

where $A_n \uparrow A$ means that

$$A_1 \subset A_2 \subset \ldots, \text{ and } A = \bigcup A_n,$$

while $A_n \downarrow A$ that

$$A_1 \supset A_2 \supset \ldots, \text{ and } A = \bigcap A_n.$$

Note, in addition, that the event Ω is sometimes said to be **certain**, and the event $\varnothing$ **impossible.**

§ 3. Independence of Events

Two events A and B are independent if $P(AB) = P(A)\cdot P(B)$.

Example 1. Consider the following probability distribution table arising in tossing a coin twice:

Ω	HH	HT	TH	TT
P	1/4	1/4	1/4	1/4

Let $A = \{HH, HT\}$ be the event "occurrence of the head in the first toss of the coin", and $B = \{HT, TT\}$ "occurrence of the tail in tossing for the second time". In this case, the events A and B are independent, since

$$AB = \{HT\}, \quad P(AB) = \frac{1}{4},$$

$$P(A) = P(HH) + P(HT) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2},$$

and

$$P(B) = P(HT) + P(TT) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

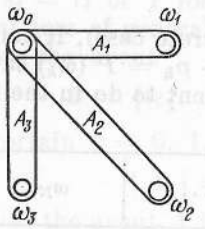

Fig. 1

Therefore, $P(AB) = P(A)\cdot P(B)$.

Let $\{A_i : i \in I\}$ be a family of events. We will say that the events of this family are **mutually independent** if

$$P\left(\bigcap_{j\in J} A_j\right) = \prod_{j\in J} P(A_j)$$

for any finite set J of indices from I.

Example 2 (*where events are pairwise independent, but dependent mutually*). Let $\Omega = \{\omega_0, \omega_1, \omega_2, \omega_3\}$, and $P(\omega_k) = 1/4$ for $k = 0, 1, 2, 3$. Consider the events $A_i = \{\omega_0, \omega_i\}$, $i = 1, 2, 3$ (see Fig. 1). As it turns out, the events A_1, A_2, A_3 are pairwise independent, but dependent mutually. In fact,

$$P(A_i) = P(\omega_0) + P(\omega_i) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \text{ for } i = 1, 2, 3,$$

and

$$P(A_iA_j) = P(\omega_0) = \frac{1}{4} = \frac{1}{2}\cdot\frac{1}{2} = P(A_i)\cdot P(A_j) \text{ for } i \neq j,$$

i.e., the events A_1, A_2, A_3 are pairwise independent, but

$$P(A_1A_2A_3) = P(\omega_0) = \frac{1}{4} \neq P(A_1)\cdot P(A_2)\cdot P(A_3) = \frac{1}{8},$$

or the events A_1, A_2, A_3 are dependent mutually.

§ 4. Conditional Probability

Consider first the classical model of probability theory (see Example 1, § 1), where

$$|\Omega| = N, \quad \text{and } P(\omega) = 1/N \quad \forall\omega \in \Omega.$$

Let A and B be two events. Assume that the event A has occurred. Then the number of elementary events has decreased from N to $n = |A|$. With this condition, denote the probability of the event B by $P_A(B)$. The event B contains $k = |AB|$ elementary events of the remaining n ones. Therefore, it is natural to define

$$P_A(B) = \frac{k}{n} = \frac{k/N}{n/N} = \frac{P(AB)}{P(A)},$$

which permits us to give the following definition of a conditional probability.

Let A and B be two events. The number $P(AB)/P(A)$, if only $P(A) \neq 0$, is called the **conditional probability** of the event B relative to A (they sometimes say: on the hypothesis that the event A has occurred), and denoted by $P_A(B)$, $P(B/A)$, or $P(B:A)$. The number $P_A(B)$ is not defined if $P(A) = 0$. Thus,

$$P(AB) = P(A)\cdot P_A(B).$$

Note that if $P(A) = 0$, then the events A and B are independent (show!). If, however, $P(A) > 0$, then the events A and B are independent if and only if $P_A(B) = P(B)$.

If (Ω, S, P) is a probability space, $A \in S$, and $P(A) > 0$, then the triplet $(\Omega_A, S_A, P_A$ is a probability space, where

$$\Omega_A = A\Omega = A, \quad S_A = \{AC : C \in S\}, \text{ and}$$

$P_A(B) = P(AB)/P(A) = P(B)/P(A)$ for any $B \in S_A$. The probability space (Ω_A, S_A, P_A) is called the **conditional probability space** relative to the event A.

§ 5. Probability of Union of Events

$$P\left(\bigcup_1^n A_i\right) = P(A_1) + \ldots + P(A_n)$$

$$(1) \quad - P(A_1A_2) - P(A_1A_3) - \ldots - P(A_{n-1}A_n) + P(A_1A_2A_3)$$
$$+ P(A_1A_2A_4) + \ldots + P(A_{n-2}A_{n-1}A_n) + (-1)^{n+1} P(A_1 \ldots A_n).$$

For $n = 2$, the formula was given in § 2. For any integer $n \geqslant 1$, it is verified by induction (on n, and using only Properties A1—A2 from § 2) on the basis of the formula

$$P\left(\bigcup_1^{n+1} A_i\right) = P\left(\bigcup_1^n A_i \cup A_{n+1}\right) = P\left(\bigcup_1^n A_i\right) + P(A_{n+1}) - P\left(\bigcup_{i=1}^n A_iA_{n+1}\right).$$

If the union operation is replaced by intersection, then the obtained formula is also valid. It can be derived from the previous one by using the relation

$$\bigcap_1^n A_i = \left(\bigcup_1^n A_i^c\right)^c,$$

and employed for computing the probability of the intersection of events (in terms of the probability of their unions) but can be proved easier also by induction (on n). Relation (1) is called the **Boole formula.**

Example (Coincidence Problem). Consider n elements placed in a certain order, and permute them at random, which means that all the $n!$ permutations are equally likely. What is the probability that at least one element is in its place?

Let A_i be the event "i-th element is in its place". There are altogether $n!$ elementary events (permutations). The event A_i contains $(n-1)!$ elementary events (permutations). Therefore,

$$P(A_i) = \frac{(n-1)!}{n!}.$$

When $i \neq j$, the event A_iA_j means that the i-th and j-th elements are in their places; hence,

$$P(A_iA_j) = \frac{(n-2)!}{n!}, \quad \ldots, \quad P(A_1 \ldots A_n) = \frac{1}{n!}.$$

The event $\bigcup_1^n A_i$ means that at least one element will be in its place. The probability of such an event is just what is required to determine. Since the addends in each line of formula (1) are the same, and there are ${}^nC_1, {}^nC_2, \ldots, {}^nC_n$ of them, respectively, we have

$$P\left(\bigcup_1^n A_i\right) = {}^nC_1 \frac{(n-1)!}{n!} - {}^nC_2 \frac{(n-2)!}{n!}$$

$$+ {}^nC_3 \frac{(n-3)!}{n!} - \ldots + (-1)^{n+1} {}^nC_n \frac{1}{n!}$$
$$= 1 - \frac{1}{2!} + \frac{1}{3!} - \ldots + (-1)^{n+1} \frac{1}{n!}$$
$$= 1 - \left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \ldots + (-1)^n \frac{1}{n!}\right).$$

The expression in the bracket is the sum of the first $n + 1$ terms of the expansion in a series of e^{-x} in terms of powers of x, and when $x = 1$. Note that

$$P\left(\bigcup_1^n A_i\right) \to 1 - e^{-1} \text{ as } n \to \infty.$$

§ 6. Composite Probability Formula

Let $\Omega = \sum_k E_k$, and $P(E_k) > 0$. Then

$$\boxed{P(A) = \sum_k P(E_k) \cdot P_{E_k}(A)},$$

which follows from the equality $A = A\Omega = \sum AE_k$, σ-additivity of probability P, and conditional probability definition.

Example. Among N examination cards, there are n "lucky". Let A_k be the event "k-th card drawn is lucky". If the cards are returned (and thoroughly mixed), then $P(A_k) = n/N$. We find $P(A_2)$ with the assumption that the cards are not replaced, and that the previous card is unknown. By the composite probability formula, we have since $(\Omega = A_1 + A_1^c = E_1 + E_2)$

$$P(A_2) = P(A_1) P(A_2 \mid A_1) + P(A_1^c) P_s(A_2 \mid A_1^c)$$
$$= \frac{n}{N} \cdot \frac{n-1}{N-1} + \frac{N-n}{N} \cdot \frac{n}{N-1} = \frac{n}{N}.$$

As it turns out, $P(A_k) = n/N$ if $1 \leqslant k \leqslant N$, and the probability $P(A_k)$ of drawing a lucky card does not depend on whether or not the previous card is replaced. Applied to the students who strive to come to an examination last, this example shows that their hopes of all unlucky question cards having already been exhausted are vain.

§ 7. Bayes Formula

Let $\Omega = \sum E_k$ again, the distribution of the events E_k be known, i.e., all $P(E_k)$ given. Suppose an event A has occurred. What is the new distribution of the events E_k on receiving this new information? More precisely, this means that it is required to determine $P_A(E_k)$ in terms of $P(E_i)$ and $P_{E_i}(A)$ $\forall i$. The answer is

given by the Bayes formula

$$P_A(E_k) = \frac{P(E_k)\,P_{E_k}(A)}{\sum_i P(E_i)\,P_{E_i}(A)},$$

which follows from the equality

$$P_A(E_k) = \frac{P(AE_k)}{P(A)}$$

for computing conditional probability, and also from the composite probability formula. It is assumed that $P(A) > 0$ and, $P(E_i) > 0\ \forall i$.

Note that the set $\{P(E_i)\}$ is often called an **a priori distribution**, and $\{P_A(E_i)\}$ **a posteriori distribution** relative to the event A (provided the event A has occurred).

§ 8. Examples of Distributions

Many classical models of probability theory are reduced to the following examples.

Example 1. Given a box containing m balls of which m_1 are white, m_2 black, and $m_1 + m_2 = m$. The balls are thoroughly mixed, one of them is taken out, its colour remembered, and then the ball is replaced. The operation is repeated n times, the balls being mixed thoroughly after each replacement. Let A_k be the event "k taken balls are white (and $n - k$ black)". Let us find the probability of this event.

If all the balls are numbered 1, 2, . . ., m, then the outcome of n draws can be represented in the form

$$\omega = (\varepsilon_1, \ldots, \varepsilon_n),$$

where ε_i denotes the number of a ball after the i-th draw. Each ε_i thereby assumes the m values 1, 2, . . ., m. We take $\Omega = \{\omega\}$ to be the space of elementary events, and put $P(\omega) = 1/|\Omega|$. The set Ω consists of $|\Omega| = m^n$ elementary events, whereas the set A_k of

$$|A_k| = {}^nC_k m_1^k m_2^{n-k}$$

elementary events. Therefore,

$$p_k = P(A_k) = \frac{|A_k|}{|\Omega|} = {}^nC_k \left(\frac{m_1}{m}\right)^k \left(\frac{m_2}{m}\right)^{n-k},$$

or

$$p_k = {}^nC_k p^k q^{n-k}, \quad k = 0, 1, \ldots, n,$$

where

$$p = \frac{m_1}{m}, \quad q = 1 - p = \frac{m_2}{m}.$$

A set of such numbers $(p_0, p_1, \ldots, p_n)$ is called the **binomial distribution** with parameters n and p.

Example 2. Consider the same model of drawing the balls, but without replacement. We find the probability $P(A_k)$ of the event A_k again, where A_k is "k balls are white among n balls taken out". It is clear that $n \leqslant m$ in this case. Number all the balls 1, 2, . . ., m again. Then the outcome of n draws is written in the form

$$\omega = (\varepsilon_1, \ldots, \varepsilon_n),$$

where ε_i denotes the number of a ball after the i-th draw, and all the numbers $\varepsilon_1, \ldots, \varepsilon_n$ are different. If $\Omega = \{\omega\}$, then

$$|\Omega| = m(m-1)\ldots(m-n+1) = m^{(n)}.$$

The event A_k consists of

$$|A_k| = {}^nC_k m_1^{(k)} m_2^{(n-k)}, \quad k \leqslant m_1$$

elementary events. Hence,

$$p_k = P(A_k) = \frac{|A_k|}{|\Omega|} = {}^nC_k \frac{m_1^{(k)} m_2^{(n-k)}}{m^{(n)}}, \quad k = 0, 1, \ldots, n.$$

The set $(p_0, p_1, \ldots, p_n)$ of these numbers is called the **hypergeometric distribution.**

Example 3. We will now assume that m balls are painted in r colours, and not in two as in the previous examples. Let m_i denote the number of balls of colour i, $i = 1, \ldots, r$, $m_1 + \ldots m_r = m$. Consider the scheme of drawing the balls and replacing them in this example. Let A_k be the event "k_1 balls are of colour 1 on the n-th draw, k_2 of colour 2, etc., k_r balls of colour r", $k = (k_1, \ldots, k_r)$, $k_1 + \ldots + k_r = n$. Then

$$p_k = P(A_k) = \frac{n!}{k_1! \ldots k_r!}\, p_1^{k_1} \ldots p_r^{k_r},$$

where $p_i = m_i/m$, and k assumes its values from $\{k = (k_1, \ldots, k_r)$: $k_1 \geqslant 0$, $k_1 + \ldots + k_r = n\}$. The set $\{p_k\}$ of these numbers is called the **multinomial distribution.**

Example 4. Now, if the balls are taken out without replacement, then

$$p_k = P(A_k) = \frac{n!}{k_1! \ldots k_r!} \cdot \frac{m_1^{(k_1)} \ldots m_r^{(k_r)}}{m^{(n)}}, \quad n \leqslant m, \quad k_i \leqslant m_i\ \forall_i.$$

The set $\{p_k\}$ of these numbers is called the **multidimensional hypergeometric distribution.**

§ 9. Random Variable. Mathematical Expectation

In probability theory, an arbitrary mapping $\xi : \Omega \to R^1$ of a space $\Omega = \{\omega_k\}$ of elementary events into the real straight line R^1 is usually called a **real random variable** (RRV). The mapping

$\xi : \Omega \to R^n$ is called a **vector-valued random variable** (VRV), R^n being an n-dimensional vector space. Recall that this definition is given only for the case where Ω is at most a countable set of elementary events.

Example 1. With the data of Example 3, § 1, we define an RRV ξ equal to the duration of the game by putting

$$\xi(\omega_k) = k, \quad k = 1, 2, \ldots$$

An RRV ξ is said to be **summable** if the series

$$\sum_k \xi(\omega_k) P(\omega_k)$$

converges absolutely, in which case the sum of the series, i.e.,

$$E\xi = \sum_{\omega \in \Omega} \xi(\omega) P(\omega),$$

is called the **mathematical expectation** (ME), or **mean value**, of the RRV.

Example 2. The average duration of the game defined in Example 3, § 1, is equal, according to Example 1 of the present section, to

$$E\xi = \sum_{k \geqslant 1} k 2^{-k} = 2.$$

In practical problems and in defining an RRV, it is essential to list the values assumed by this variable, and also the corresponding probabilities. In this sense, an RRV ξ can conveniently be specified by means of the table of the distribution of RRV values, viz.,

$$\frac{\xi}{p} \left| \begin{array}{c|c|c|c|c} x_1 & x_2 & \ldots & x_N & \ldots \\ \hline p_1 & p_2 & \ldots & p_N & \ldots \end{array} \right.,$$

where

$$p_k = P(\xi = x_k), \quad \text{and} \quad \sum_{k \geqslant 1} p_k = 1.$$

Then

$$E\xi = \sum_{k \geqslant 1} x_k p_k,$$

which is justified by the following lemma.

Lemma. Let an RRV ξ assume some values $x_1, x_2, \ldots$ on pairwise disjoint sets $A_1, A_2, \ldots$. Then, for the summability of the RRV, the absolute convergence of the series

$$\sum_k x_k P(A_k)$$

is necessary and sufficient. Besides,

$$E\xi = \sum_k x_k P(A_k)$$

Proof. If one of the series

$$\sum_\omega |\xi(\omega)| P(\omega) \quad \text{and} \quad \sum_k |x_k| P(A_k)$$

converges, then the other is also convergent, which follows from the relation

$$\sum_{\omega \in \Omega} |\xi(\omega)| P(\omega) = \sum_k \Big[\sum_{\omega \in A_k} |\xi(\omega)| P(\omega) \Big]$$

$$= \sum_k |x_k| \sum_{\omega \in A_k} P(\omega) = \sum_k |x_k| P(A_k).$$

It is verified similarly that

$$E\xi = \sum_\omega \xi(\omega) P(\omega) = \sum_k x_k P(A_k).$$

Remark. In defining an ME, the **absolute** convergence of the corresponding series is required. Otherwise, the lemma is not valid. If, however, the series converges absolutely, then regrouping the addends of the series without altering the sum is permitted, which is just what we have used in the proof of the lemma.

Example 3. With the data of Example 1, § 5, let ξ be the number of the white balls taken out. It has been shown that

$$P(\xi = k) = {}^nC_k p^k q^{n-k}, \quad k = 0, 1, \ldots, n.$$

Note that $\{\xi = k\}$ is the contracted form of the notation of the event

$$\{\omega : \xi(\omega) = k\} = A_k.$$

The average number of the white balls taken out is

$$E\xi = \sum_{k=0}^{n} k \, {}^nC_k p^k q^{n-k} = np.$$

§ 10. Independence of Random Variables

Let ξ and η be two RRV's assuming values only from two sets $\{x_i\}$ and $\{y_j\}$, respectively. By the definition of an RRV, ξ and η are independent if the events

$$\{\xi = x_i\} \quad \text{and} \quad \{\eta = y_j\}$$

are independent for any pair (i, j). The mutual independence of a sequence of RRV's is defined similarly (in terms of the mutual independence of events, see § 3).

Lemma 1. If $\{\xi_1, \ldots, \xi_m, \eta_1, \ldots, \eta_n\}$ are mutually independent RRV's, then the RRV's

$$f(\xi_1, \ldots, \xi_m) \quad \text{and} \quad g(\eta_1, \ldots, \eta_n)$$

are also independent (f and g are arbitrary real functions).

Proof. Let

$$\xi = (\xi_1, \ldots, \xi_m), \quad \eta = (\eta_1, \ldots, \eta_n).$$

It follows from the conditions of the lemma that

$$P\{\xi = x,\ \eta = y\} = P\{\xi = x\}\cdot P\{\eta = y\}$$

for arbitrary

$$x = (x_1, \ldots, x_m), \qquad y = (y_1, \ldots, y_n).$$

It should be shown that

$$P_{\cdot}\{f(\xi) = u,\ \ g(\eta) = v\} = P\{f(\xi) = u\}\cdot P\{g(\eta) = v\}$$

for arbitrary numbers u and v. In fact,

$$\begin{aligned} P\{f(\xi)=u,\ g(\eta)=v\} &= P\{\xi\in f^{-1}(u),\ \ \eta\in g^{-1}(v)\} \\ &= P\{\xi\in f^{-1}(u)\}\cdot P\{\eta\in g^{-1}(v)\} \\ &= P\{f(\xi)=u\}\cdot P\{g(\eta)=v\}, \end{aligned}$$

since

$$\begin{aligned} P\{\xi\in A,\ \eta\in B\} &= P\Big\{\sum_{x\in A}\sum_{y\in B}(\xi=x,\ \eta=y)\Big\} \\ &= \sum_{x,\,y} P(\xi=x,\ \eta=y) = \sum_{x,\,y} P(\xi=x)\cdot P(\eta=y) \\ &= \sum_{x\in A} P(\xi=x)\cdot\sum_{y\in B} P(\eta=y) = P\Big\{\sum_{x\in A}(\xi=x)\Big\}\cdot P\Big\{\sum_{y\in B}(\eta=y)\Big\} \\ &= P\{\xi\in A\}\cdot P\{\eta\in B\}. \end{aligned}$$

Lemma 2. If $\{\xi_1, \ldots, \xi_n\}$ are mutually independent RRV's, then the RRV's

$$f_1(\xi_{i_1}, \ldots, \xi_{j_1}),\ \ f_2(\xi_{i_2}, \ldots, \xi_{j_2}),\ \ f_m(\xi_{i_m}, \ldots, \xi_{j_m})$$

are also mutually independent (where $1 \leqslant i_1 \leqslant j_1 < i_2 \leqslant j_2 < \ldots$ $\ldots < i_m \leqslant j_m \leqslant n$; $f_1, \ldots, f_m$ are arbitrary real functions).

Proof can be performed by the same method as the one employed for Lemma 1 or by the reduction to Lemma 1.

§ 11. Additive and Multiplicative Properties of Mathematical Expectation

1°. **Additivity of an ME.** Let ξ and η be two summable RRV's, and λ an arbitrary number. Then $\xi + \eta$ and $\lambda\xi$ are summable RRV's, while

$$E(\xi + \eta) = E\xi + E\eta, \quad E(\lambda\xi) = \lambda E\xi.$$

This follows from the properties of absolutely convergent series.

2°. **Multiplicativity of an ME.** Let ξ and η be two summable and independent RRV's, then the RRV $\xi\cdot\eta$ is summable, whereas

$$E(\xi\cdot\eta) = E\xi\cdot E\eta$$

(if the independence of the RRV's ξ and η is not required, then, to prove the summability of the RRV $\xi\cdot\eta$, the separate summability of ξ and η is insufficient).

Proof. Let $\{x_i\}$ be the collection of all values of the RRV ξ, and $\{y_j\}$ that of η. Put

$$A_i = \{\omega : \xi(\omega) = x_i\}, \quad B_j = \{\omega : \eta(\omega) = y_j\}.$$

Due to the independence of the RRV's ξ and η, we have

$$P(A_iB_j) = P(A_i)\cdot P(B_j).$$

Now,

$$\begin{aligned} E(\xi\cdot\eta) &= \sum_{\omega}\xi(\omega)\,\eta(\omega)\,P(\omega) = \sum_{i,\,j}\sum_{\omega\in A_iB_j}\xi(\omega)\,\eta(\omega)\,P(\omega) \\ &= \sum_{i,\,j} x_iy_j\sum_{\omega\in A_iB_j} P(\omega) = \sum_{i,\,j} x_iy_jP(A_iB_j) = \sum_{i,\,j} x_iy_jP(A_i)\cdot P(B_j) \\ &= \sum_i x_iP(A_i)\cdot\sum_j y_jP(B_j) = E\xi\cdot E\eta, \end{aligned}$$

since, in view of the summability of the RRV's ξ and η, it follows that the latter two series are absolutely convergent, thus justifying their multiplying together, and resulting in the absolute convergence of the obtained series. Besides, we have used here the Lemma, § 9.

Corollary. Let $\xi_1, \ldots, \xi_n$ be summable and independent RRV's, then the RRV $\xi_1, \ldots \xi_n$ is summable, with

$$E(\xi_1 \ldots \xi_n) = E\xi_1 \ldots E\xi_n.$$

Proof is performed by induction, and based on Lemma 1, § 10. In fact, for $n = 2$, the statement has already been proved. By Lemma 1, § 10, the RRV's $(\xi_1 \ldots \xi_{n-1})$ and ξ_n are independent; hence,

$$E(\xi_1 \ldots \xi_n) = E(\xi_1 \ldots \xi_{n-1})\cdot\xi_n = E(\xi_1 \ldots \xi_{n-1})\cdot E\xi_n.$$

§ 12. Moments

Let ξ be an RRV. The number $E\xi^k$ is called the ***k*-th moment** (if the RRV ξ^k is not summable, then we put $E\xi^k = \infty$).

The number $E(\xi - E\xi)^k$ is called a **centred *k*-th moment**. In particular, when $k = 2$, we obtain the number $\operatorname{var}\xi = D\xi = E(\xi - E\xi)^2$ called the variance of the RRV ξ.

For any pair of RRV's ξ and η, the numbers

$$\operatorname{cov}(\xi,\ \eta) = E(\xi - E\xi)(\eta - E\eta), \quad \operatorname{cor}(\xi,\ \eta) = \frac{\operatorname{cov}(\xi,\ \eta)}{(\operatorname{var}\xi\cdot\operatorname{var}\eta)^{1/2}},$$

are called the **covariance** and **correlation** of the RRV's ξ and η. If $\operatorname{var}\xi = 0$ and/or $\operatorname{var}\eta = 0$, then $\operatorname{cor}(\xi,\ \eta)$ is not defined.

Note that always (show!)

$$-1 \leqslant \operatorname{cor}(\xi,\ \eta) \leqslant +1,$$

while the independence of the RRV's ξ and η is sufficient for $\operatorname{cor}(\xi, \eta) = 0$. If $\operatorname{cor}(\xi, \eta) = 0$ then the RRV's ξ and η are said to be **uncorrelated**. In practice, if some RRV's are uncorrelated, then it is a serious argument for their independence (though this is not always true).

§ 13. Variance Additivity for Independent RRV's

Let $\xi_1 \ldots \xi_n$ be n independent RRV's (it is sufficient that they should be pairwise uncorrelated), and $\xi = \xi_1 + \ldots + \xi_n$, $E\xi_i^2 < \infty$ $\forall i$. Then

$$D\xi = D\xi_1 + \ldots + D\xi_n,$$

which follows from the definition of variance, and also from the multiplicativity of the ME for independent RRV's (see § 11).

Remark. If, in addition, $E\xi_i = 0$, $1 \leqslant i \leqslant n$, then $E\xi^3 = \sum E\xi_i^3$, but $E\xi^4 = \sum E\xi_i^4 + 3 \sum_{i \neq j} E\xi_i^2 \cdot E\xi_j^2 \neq \sum E\xi_i^4$.

§ 14. Generating Function

Let an RRV ξ assume values $x_0, x_1, \ldots$ with probabilities $p_0, p_1, \ldots$. Associate the probability distribution $\{p_k\}$, i.e., set of numbers $(p_0, p_1, \ldots)$ such that $p_k \geqslant 0$ for $\forall k$ and $\sum_{k \geqslant 0} p_k = 1$, the complex function $p(z)$ of a complex argument z being defined in the disk $|z| \leqslant 1$ by the power series

$$p(z) = \sum_{k \geqslant 0} p_k z^k.$$

Such a function $p(z)$ is called the **generating function** of the number set $\{p_k\}$ or distribution $\{p_k\}$. The distribution can be restored uniquely from a generating function, e.g., by the formula

$$p_k = \frac{1}{2\pi i} \int_{|z|=1} \frac{p(z)}{z^{k+1}}\, dz, \quad k = 0, 1, 2, \ldots, \tag{1}$$

or by the formula

$$p_k = \frac{1}{k!} p^{(k)}(0).$$

Thus, the mapping

$$\{p_k\} \to p(z), \quad |z| \leqslant 1,$$

is one-to-one.

Example 1. Let $x_k = k$ $\forall k$. Then

$$p(z) = Ez^{\xi},$$

$$E\xi = p'(1), \quad E\xi^2 = p''(1) + p'(1),$$

$$E\xi^{(k)} = p^{(k)}(1), \quad k = 1, 2, \ldots,$$

where $\xi^{(k)} = \xi(\xi - 1) \ldots (\xi - k + 1)$.

Example 2. Let $\xi_1, \ldots, \xi_n$ be n independent RRV's assuming values from the set $N = \{0, 1, 2, \ldots\}$, and $p_1(z), \ldots, p_n(z)$ the corresponding generating functions, where $\xi = \xi_1 + \ldots + \xi_n$. Then the generating function corresponding to the RRV ξ equals

$$p(z) = p_1(z) \ldots p_n(z) \quad \text{when} \quad |z| \leqslant 1.$$

In fact, due to the multiplicativity of the ME of independent RRV's (see § 11), we have

$$p(z) = Ez^{\xi} = Ez^{\xi_1} \ldots z^{\xi_n} = Ez^{\xi_1} \ldots Ez^{\xi_n}$$
$$= p_1(z) \ldots p_n(z).$$

§ 15. Binomial Distribution

(a) Let ξ be an RRV assuming values 1 or 0 with probabilities p and $q = 1 - p$, respectively. Then

$$E\xi = 1 \cdot p + 0 \cdot q = p,$$

$$\operatorname{var} \xi = E(\xi - p)^2 = (1 - p)^2 \cdot p + (0 - p)^2 \cdot q = pq,$$

$$p(z) = Ez^{\xi} = pz + q.$$

(b) Let $\xi_1, \ldots, \xi_n$ be n independent RRV's assuming values 1 or 0 with probabilities p and $q = 1 - p$, respectively. The RRV distribution

$$\xi = \xi_1 + \ldots + \xi_n$$

is said to be **binomial**, and denoted by Bi (n, p), where

$$E\xi = np, \quad \operatorname{var} \xi = npq, \quad p(z) = Ez^{\xi} = (pz + q)^n.$$

The RRV ξ is often interpreted as the number of successes in a series of n independent trials with constant probability p of success.

(c) Let an RRV ξ only assume values $0, 1, \ldots, n$. Put

$$p_k = P(\xi = k).$$

We define p_k by the relation

$$p(z) = Ez^{\xi} = \sum_{k=0}^{n} p_k z^k = (pz + q)^n = \sum_{k=0}^{n} {}^nC_k p^k q^{n-k} z^k,$$

whence

$$p_k = {}^nC_k p^k q^{n-k}, \quad k = 0, 1, \ldots, n.$$

(d) Let ξ_1 and ξ_2 be two independent RRV's with distributions Bi (n_1, p) and Bi (n_2, p), respectively. Then the RRV $\xi = \xi_1 + \xi_2$ has the distribution Bi $(n_1 + n_2, p)$, which follows from the equalities

$$p(z) = Ez^{\xi} = Ez^{\xi_1} \cdot Ez^{\xi_2} = p_1(z) \cdot p_2(z)$$
$$= (pz + q)^{n_1} \cdot (pz + q)^{n_2} = (pz + q)^{n_1 + n_2}$$

and also from the distributions and generating functions being in one-to-one correspondence.

§ 16. Poisson Distribution

(a) Let an RRV ξ only assume values 0, 1, 2, . . ., with

$$P(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0, 1, 2, \ldots$$

for a certain $\lambda > 0$. Such a distribution is called the **Poisson distribution** with parameter λ, and denoted by Po (λ). Meanwhile,

$$p(z) = Ez^{\xi} = \sum_{k \geqslant 0} z^k \frac{\lambda^k}{k!} e^{-\lambda} = e^{\lambda(z-1)},$$

$$E\xi = p'(1) = \lambda. \quad \operatorname{var} \xi = p''(1) + p'(1) - [p'(1)]^2 = \lambda.$$

(b) Let ξ_1 and ξ_2 be two independent RRV's having the Poisson distributions with parameters λ_1 and λ_2, respectively. Then the RRV $\xi = \xi_1 + \xi_2$ also has the Poisson distribution with parameter $\lambda = \lambda_1 + \lambda_2$, which follows from the equality

$$p(z) = Ez^{\xi} = Ez^{\xi_1} \cdot Ez^{\xi_2} = e^{\lambda_1(z-1)} e^{\lambda_2(z-1)} = e^{\lambda(z-1)}.$$

§ 17. Negative Binomial Distribution

(a) Let independent trials with constant probability p of success be performed until a successful trial occurs. Suppose that the number of performed trials is $1 + \xi$. The RRV ξ can be equal only to 0, 1, 2, ... Meanwhile,

$$P(\xi = k) = q^k p, \quad k \geqslant 0, \quad q = 1 - p,$$

$$p(z) = Ez^{\xi} = \frac{p}{1 - qz}, \quad E\xi = \frac{q}{p}, \quad \operatorname{var} \xi = \frac{q}{p^2}.$$

We denote the distribution of the RRV ξ by $\overline{\text{Bi}}(1, p)$.

Remark. We can assume that the space Ω of elementary events consists of

$$\omega_0 = \text{s}, \ \omega_1 = \text{fs}, \ \omega_2 = \text{ffs}, \ \ldots, \ \omega_k = \underbrace{\text{f} \ldots \text{fs}}_{k \text{ letters}}, \ \ldots$$

(f = failure, s = success),

$$P(\omega_k) = q^k p, \quad \xi(\omega_k) = k.$$

(b) Let the same trials be performed until n successes occur. Denote the number of performed trials by $n + \xi$, $\xi = 0, 1, 2, \ldots$. We can assume that

$$\xi = \xi_1 + \ldots + \xi_n,$$

where the RRV's $\xi_1, \ldots, \xi_n$ are independent, and each has the distribution $\overline{\text{Bi}}(1, p)$. Therefore, the generating function of the RRV ξ equals

$$p(z) = \left(\frac{p}{1 - qz}\right)^n.$$

Denote the distribution of such an RRV by $\overline{\text{Bi}}(n, p)$. It is called the **negative binomial distribution**. In addition,

$$E\xi = n \cdot \frac{q}{p}, \quad \operatorname{var} \xi = n \cdot \frac{q}{p^2}.$$

(c) Find $p_k = P(\xi = k)$, $k = 0, 1, 2, \ldots$. Differentiating the left-hand and right-hand sides of the expression

$$\frac{1}{1 - z} = \sum_{N \geqslant 0} z^N$$

m times, we obtain

$$\frac{m!}{(1-z)^{m+1}} = \sum_{N \geqslant m} N(N-1) \ldots (N-m+1) z^{N-m}$$

or

$$\frac{z^m}{(1-z)^{m+1}} = \sum_{N \geqslant m} {}^{N}C_m z^N, \quad |z| < 1.$$

Using the latter identity, we get

$$p(z) = \frac{p^n}{(1-qz)^n} = \frac{p^n}{(qz)^{n-1}} \frac{(qz)^{n-1}}{(1-qz)^n} = p^n \sum_{N \geqslant n-1} {}^{N}C_{n-1} (qz)^{N-n+1}$$

$$= p^n \sum_{k \geqslant 0}^{k+n-1} C_{n-1} q^k z^n = \sum_{k \geqslant 0} p_k z^k,$$

where we have put $N - n + 1 = k$. Hence,

$$p_k = {}^{k+n-1}C_{n-1} p^n q^k; \ k = 0, 1, 2, \ldots$$

(d) If ξ_1 and ξ_2 are two independent RRV's with the distributions $\overline{\text{Bi}}(n_1, p)$ and $\overline{\text{Bi}}(n_2, p)$, respectively, then the RRV $\xi = \xi_1 + \xi_2$ has the negative binomial distribution $\overline{\text{Bi}}(n_1 + n_2, p)$, which follows from the equalities

$$p(z) = Ez^{\xi} = Ez^{\xi_1} Ez^{\xi_2} = \left(\frac{p}{1-qz}\right)^{n_1} \left(\frac{p}{1-qz}\right)^{n_2} = \left(\frac{p}{1-qz}\right)^{n_1+n_2}.$$

§ 18. Approximation of Binomial by Poisson Distribution

If, in a binomial distribution Bi (n, p), the numbers n and p vary so that

$$n \to \infty, \ np = \lambda = \text{const},$$

then

$${}^{n}C_k p^k q^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots,$$

which can be proved by the Stirling formula. A stronger statement is the following

$$\left|{}^nC_k p^k q^{n-k} - \frac{\lambda^k}{k!} e^{-\lambda}\right| \leqslant 2np^2, \quad \lambda = np, \quad p \leqslant \frac{1}{2},$$

the inequality being held for all $k = 0, 1, 2, \ldots$ if we assume that ${}^nC_k p^k q^{n-k} = 0$ for $k > n$.

We now prove a more general statement.

Let $\xi_1, \ldots, \xi_n$ be n independent RRV's assuming values 1 or 0, while

$$P(\xi_s = 1) = p_s, \; P(\xi_s = 0) = q_s = 1 - p_s;$$

$$s = 1, \ldots, n.$$

Put

$$\xi = \xi_1 + \ldots + \xi_n$$

$$\alpha_k = P(\xi = k), \quad \beta_k = \frac{\lambda^k}{k!} e^{-\lambda}, \; k = 0, 1, 2, \ldots,$$

$$\lambda = p_1 + \ldots + p_n, \quad \delta = p_1^2 + \ldots + p_n^2.$$

Theorem. Let $p_s \leqslant 1/2$ for $\forall s$. Then

$$|\alpha_k - \beta_k| \leqslant 2\delta \quad \text{for} \quad \forall k.$$

Proof. (a) The generating functions of the distributions $\{\alpha_k\}$ and $\{\beta_k\}$ are equal, respectively, to

$$p_1(z) = Ez^{\xi} = Ez^{\xi_1} \ldots Ez^{\xi_n} = \prod_{s=1}^{n} (p_s z + q_s)$$

and

$$p_2(z) = \sum_{k \geqslant 0} \beta_k z^k = e^{\lambda(z-1)}.$$

Since (see (1) in § 14)

$$\alpha_k = \frac{1}{2\pi i} \int_{|z|=1} \frac{p_1(z)}{z^{k+1}}\, dz, \quad \beta_k = \frac{1}{2\pi i} \int_{|z|=1} \frac{p_2(z)}{z^{k+1}}\, dz,$$

we have

$$(1) \quad |\alpha_k - \beta_k| = \left| \frac{1}{2\pi i} \int_{|z|=1} \frac{p_1(z) - p_2(z)}{z^{k+1}} dz \right|$$

$$\leqslant \frac{1}{2\pi} \int_0^{2\pi} |p_1(e^{i\varphi}) - p_2(e^{i\varphi})|\, d\varphi.$$

To estimate the right-hand side of this formula, we make use of the following two statements.

(b) Let $a_1, \ldots, a_n, b_1, \ldots, b_n$ be complex numbers such that $|a_s| \leqslant 1$ and $|b_s| \leqslant 1$ for $s = 1, \ldots, n$. Then

$$\left|\prod_1^n a_s - \prod_1^n b_s\right| \leqslant \sum_1^n |a_s - b_s|.$$

This inequality is verified by induction. When $n = 1$, it is trivial. In the case $n = 2$, we have

$$|a_1 a_2 - b_1 b_2| = |(a_1 a_2 - b_1 a_2) + (b_1 a_2 - b_1 b_2)$$

$$\leqslant |a_1 - b_1| \cdot |a_2| + |a_2 - b_2| \cdot |b_1|$$

$$\leqslant |a_1 - b_1| + |a_2 - b_2|.$$

Using the inequality already verified for the case $n = 2$, and assuming that it is held for $n = m - 1$, we obtain

$$\left|\prod_1^m a_s - \prod_1^m b_s\right| = \left|a_m \cdot \prod_1^{m-1} a_s - b_m \cdot \prod_1^{m-1} b_s\right|$$

$$\leqslant |a_m - b_m| + \left|\prod_1^{m-1} a_s - \prod_1^{m-1} b_s\right|$$

$$\leqslant |a_m - b_m| + \sum_1^{m-1} |a_s - b_s| = \sum_1^m |a_s - b_s|.$$

(c) If w is a complex number and $|w| \leqslant 1$, then

$$|e^w - 1 - w| \leqslant |w|^2.$$

In fact,

$$|e^w - 1 - w| = \left|\frac{w^2}{2!} + \frac{w^3}{3!} + \ldots\right| \leqslant |w|^2 \left(\frac{1}{2!} + \frac{1}{3!} + \ldots\right)$$

$$= |w|^2 (e - 2) \leqslant |w|^2.$$

(d) Putting

$$a_s = p_s z + q_s = 1 + p_s (z - 1), \quad b_s = e^{p_s(z-1)},$$

we obtain for $|z| \leqslant 1$, that $|a_s| \leqslant 1$, and $|b_s| \leqslant 1$. Therefore,

$$|p_1(z) - p_2(z)| = \left|\prod_1^n a_s - \prod_1^n b_s\right| \leqslant \sum_1^n |a_s - b_s|$$

$$= \sum_1^n |e^{w_s} - 1 - w_s| \leqslant \sum_1^n |w_s|^2$$

$$= \delta |z - 1|^2 = 2\delta (1 - \cos\varphi),$$

where $w_s = p_s(z-1)$ and $z = e^{i\varphi}$. But then the right-hand side of equality (1) is majorized by the number

$$\frac{1}{2\pi}\int_0^{2\pi} 2\delta(1-\cos\varphi)\,d\varphi = 2\delta.$$

§ 19. Indicator of Event

Let A be an event. A function I_A defined by the relation

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \overline{\in} A \end{cases}$$

is called the **indicator** of the event A. In analysis, such a function is called a **characteristic function** (CF). In probability theory and statistics, the latter term is used for other functions (see § 10, Ch. 2). Note the basic properties of indicators, viz.,

1°. I_A is a summable RRV,

2°. $EI_A = P(A)$,

and

3°. $I_A^k = I_A$ for $k = 1, 2, \ldots$, $I_{A^c} = 1 - I_A$, $I_{AB} = I_A \cdot I_B$, $I_{A+B} = I_A + I_B$.

Example (*of the use of indicators*). Since $A \cup B = A + BA^c$,

$$I_{A \cup B} = I_A + I_{BA^c} = I_A + I_B \cdot I_{A^c} = I_A + I_B - I_{AB},$$

whence

$$P(A \cup B) = EI_{A \cup B} = P(A) + P(B) - P(AB).$$

§ 20. Markov Inequality

For any $k \geqslant 0$ and $\varepsilon > 0$,

$$\boxed{P\{|\xi| \geqslant \varepsilon\} \leqslant \frac{E|\xi|^k}{\varepsilon^k}}.$$

In fact, if $A = \{|\xi| \geqslant \varepsilon\}$, then

$$|\xi|^k \geqslant |\xi|^k \cdot I_A \geqslant \varepsilon^k I_A,$$

whence

$$E|\xi|^k \geqslant \varepsilon^k EI_A = \varepsilon^k P(A).$$

Corollary. Putting $k = 2$, and replacing ξ by $\xi - E\xi$, we obtain the **Chebyshev inequality**

$$P\{|\xi - E\xi| \geqslant \varepsilon\} \leqslant \frac{\operatorname{var}\xi}{\varepsilon^2}.$$

§ 21. Law of Large Numbers

For each n, we define mutually independent RRV's $\xi_1, \ldots, \xi_n$ assuming values 1 or 0 with probabilities p and $q = 1 - p$, respectively. Put $S_n = \xi_1 + \ldots + \xi_n$. In this case, n independent trials with constant probability p of success are often spoken of. Then S_n is the number of successes in n trials.

Remark 1. The corresponding space Ω of elementary events can be represented in the form

$$\Omega = \{\omega\}, \quad \omega = (\varepsilon_1, \ldots, \varepsilon_n),$$

where ε_i assume values 1 or 0, with $|\Omega| = 2^n$, $P(\omega) = p^k q^{n-k}$, and $k = k(\omega) = \varepsilon_1 + \ldots + \varepsilon_n$, $\xi_k(\omega) = \varepsilon_k$.

The following statement is the simplest version of the **Law of Large Numbers (LLN).**

Theorem. For any $\varepsilon > 0$,

$$P\left\{\left|\frac{S_n}{n} - p\right| \geqslant \varepsilon\right\} \to 0 \text{ as } n \to \infty.$$

Remark 2. It follows from this assertion that the share S_n/n of successes in n trials approaches the probability of one successful trial as $n \to \infty$ (frequency tends to probability!).

Proof immediately follows from the Chebyshev inequality (see § 20):

$$P\left\{\left|\frac{S_n}{n} - p\right| \geqslant \varepsilon\right\} \leqslant \frac{\operatorname{var}\frac{S_n}{n}}{\varepsilon^2} = \frac{pq}{n\varepsilon^2}, \tag{1}$$

Corollary 1. Let $f(x)$ be a continuous real function on $[0, 1]$. Then, as $n \to \infty$ uniformly with respect to $p \in [0, 1]$,

$$Ef\left(\frac{S_n}{n}\right) \to f(p).$$

In fact, putting

$$A_n = \left\{\left|\frac{S_n}{n} - p\right| \geqslant \varepsilon\right\}$$

for $\varepsilon > 0$, we obtain, making use of formula (1),

$$\begin{aligned}\left|Ef\left(\frac{S_n}{n}\right) - f(p)\right| &\leqslant E\left|f\left(\frac{S_n}{n}\right) - f(p)\right| \\ &= E\left|f\left(\frac{S_n}{n}\right) - f(p)\right| \cdot I_{A_n^c} \\ &\quad + E\left|f\left(\frac{S_n}{n}\right) - f(p)\right| \cdot I_{A_n} \\ &\leqslant \sup_{|x|<\varepsilon} |f(p+x) - f(p)| + 2c \cdot EI_{A_n} \\ &\leqslant \sup_{|x|<\varepsilon} |f(p+x) - f(p)| + 2c \cdot \frac{pq}{n\varepsilon^2},\end{aligned}$$

where c is determined by the condition $|f(x)| \leqslant c\ \forall x$. It remains to notice that the function f is uniformly continuous on the interval $[0, 1]$.

Corollary 2 (Weierstrass Approximation Theorem). Let $f(x)$ be a continuous real function on $[0, 1]$. Then, as $n \to \infty$ uniformly with respect to $x \in [0, 1]$,

$$P_n(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) {}^nC_k x^k (1-x)^{n-k} \to f(x).$$

This relation is just another way of stating Corollary 1, since

$$P(S_n = k) = {}^nC_k p^k (1-p)^{n-k}.$$

The latter statement is the Weierstrass theorem on uniform approximation by polynomials of a function continuous on an interval. Here, these polynomials $P_n(x)$ are written out explicity (and called **Bernstein polynomials**).

Khinchine Law of Large Numbers. Let $\xi_1, \xi_2, \ldots$ be independent and identically distributed (i.i.d.) RRV's and $E|\xi_1| < \infty$. Put

$$S_n = \xi_1 + \ldots + \xi_n,\ E\xi_1 = \mu.$$

Then

$$\frac{S_n}{n} \to \mu$$

in probability, i.e., for any $\varepsilon > 0$ as $n \to \infty$,

$$P\left\{\left|\frac{S_n}{n} - \mu\right| \geqslant \varepsilon\right\} \to 0.$$

Proof (*which remains unaltered also in the case of a general probability space*).

1°. Use the following statements: if, for an RV ξ, the inequality $E|\xi| < \infty$ holds, then

$$nP(|\xi| \geqslant n) \to 0, \tag{2}$$

$$\frac{1}{n} E|\xi|^2 \cdot 1_{\{|\xi|<n\}} \to 0 \text{ as } n \to \infty. \tag{3}$$

Let us prove (2). $\forall \varepsilon > 0 \exists N$ such that

$$E|\xi| \cdot 1_{\{|\xi| \geqslant n\}} \leqslant \varepsilon \text{ for } n \geqslant N;$$

whence, taking into account that

$$nE1_{\{|\xi|\geqslant n\}} \leqslant E|\xi| \cdot 1_{\{|\xi|\geqslant n\}},$$

we obtain

$$nP(|\xi| \geqslant n) \leqslant \varepsilon \text{ for } n \geqslant N.$$

We now derive (3) from (2). Putting $q_n = P(|\xi| \geqslant n)$, we have

$$\begin{aligned} E|\xi|^2 1_{\{|\xi|<n\}} &= \sum_1^n E|\xi|^2 1_{\{k-1\leqslant|\xi|<k\}} \\ &\leqslant \sum_1^n k^2 E\, 1_{\{k-1\leqslant|\xi|<k\}} = \sum_1^n k^2 (q_{k-1} - q_k) \\ &= q_0 - n^2 q_n + \sum_1^{n-1} q_k [(k+1)^2 - k^2] \\ &\leqslant q_0 + \sum_{k=1}^{n-1} (2k+1)\, q_k. \end{aligned}$$

According to (2), for any $\varepsilon > 0$, there exists N such that $kq_k \leqslant \varepsilon$ when $k \geqslant N$; whence

$$\begin{aligned} \frac{1}{n} E|\xi|^2 1_{\{|\xi|<n\}} &\leqslant \frac{1}{n} q_0 + \frac{1}{n} \sum_1^N (2k+1)\, q_k + \frac{1}{n} \sum_{N<k<n} 3kq_k \\ &\leqslant \frac{1}{n} q_0 + \frac{1}{n} \sum_1^N (2k+1)\, q_k + 3\varepsilon \text{ for } n > N, \end{aligned}$$

which gives (3).

2°. For each integer $n \geqslant 1$, we form RV S_n^* by putting

$$S_n^* = \xi_1^* + \ldots + \xi_n^*,\ \xi_k^* = \begin{cases} \xi_k, & |\xi_k| < n \\ 0, & |\xi_k| \geqslant n \end{cases}$$

and show that

$$\frac{S_n^*}{n} \to \mu$$

in probability. In fact, it follows from the independence of the RV's $\xi_1^*, \ldots, \xi_n^*$ that

$$\begin{aligned} P\left\{\left|\frac{S_n^*}{n} - \mu\right| \geqslant \varepsilon\right\} &\leqslant \frac{1}{\varepsilon^2} E\left(\frac{S_n^*}{n} - \mu\right)^2 \\ &= \frac{1}{\varepsilon^2}\left[\operatorname{var}\frac{S_n^*}{n} + \left(E\frac{S_n^*}{n} - \mu\right)^2\right] \\ &= \frac{1}{\varepsilon^2}\left[\frac{1}{n}\operatorname{var}\xi_1^* + (E\xi_1^* - \mu)^2\right]. \end{aligned}$$

It remains to notice that, because of (3) and $E|\xi_1| < \infty$,

$$\begin{aligned} \frac{1}{n}\operatorname{var}\xi_1^* &\leqslant \frac{1}{n} E|\xi_1^*|^2 = \frac{1}{n} E\xi_1^2 \cdot 1_{\{|\xi_1|<n\}} \to 0, \\ E\xi_1^* &= E\xi_1 1_{\{|\xi_1|<n\}} \to E\xi_1 = \mu. \end{aligned}$$

3°. To show that

$$\frac{S_n}{n} - \frac{S_n^*}{n} \to 0$$

in probability, we derive from the Markov inequality for $\varepsilon > 0$ that

$$P\left\{\left|\frac{S_n}{n}-\frac{S_n^*}{n}\right|\geqslant\varepsilon\right\}\leqslant\frac{1}{\varepsilon}E\left|\frac{S_n}{n}-\frac{S_n^*}{n}\right|$$

$$\leqslant\frac{1}{\varepsilon n}\sum_1^n E\,|\xi_k-\xi_k^*|=\frac{1}{\varepsilon}E\,|\xi_1|\cdot 1_{\{|\xi_1|\geqslant n\}}\to 0.$$

4°. From 2° and 3°, we obtain, finally, that

$$P\left\{\left|\frac{S_n}{n}-\mu\right|\geqslant\varepsilon\right\}\leqslant P\left\{\left|\frac{S_n^*}{n}-\mu\right|+\left|\frac{S_n}{n}-\frac{S_n^*}{n}\right|\geqslant\varepsilon\right\}$$

$$\leqslant P\left\{\left|\frac{S_n^*}{n}-\mu\right|\geqslant\frac{\varepsilon}{2}\ \text{ or }\ \left|\frac{S_n}{n}-\frac{S_n^*}{n}\right|\geqslant\frac{\varepsilon}{2}\right\}$$

$$\leqslant P\left\{\left|\frac{S_n^*}{n}-\mu\right|\geqslant\frac{\varepsilon}{2}\right\}+P\left\{\left|\frac{S_n}{n}-\frac{S_n^*}{n}\right|\geqslant\frac{\varepsilon}{2}\right\}\to 0.$$

Remark. A.Ya. Khinchine proved that, in the case of i.i.d. RV's $\xi_1, \xi_2, \ldots$, for $\frac{S_n}{n}=\frac{1}{n}(\xi_1+\ldots+\xi_n)$ to be convergent to a certain number μ in probability, it is necessary and sufficient that

$$nP(|\xi_1|\geqslant n)\to 0,\quad E\xi_1\cdot 1_{\{|\xi_1|<n\}}\to\mu.$$

The sufficiency of these conditions will be, in fact, proved if the end of Item 3° of the proof (where we have made use of the inequality $E|\xi_1|<\infty$) is replaced by the following argument, viz.,

$$P\left\{\left|\frac{S_n}{n}-\frac{S_n^*}{n}\right|\geqslant\varepsilon\right\}\leqslant P\left(\frac{S_n}{n}-\frac{S_n^*}{n}\neq 0\right)$$

$$=1-P\left(\frac{S_n}{n}-\frac{S_n^*}{n}=0\right),$$

and, since

$$\left\{\frac{S_n}{n}-\frac{S_n^*}{n}=0\right\}\supset\{|\xi_1|<n,\ \ldots,\ |\xi_n|<n\},$$

we obtain

$$P\left\{\left|\frac{S_n}{n}-\frac{S_n^*}{n}\right|\geqslant\varepsilon\right\}\leqslant 1-P\{|\xi_k|<n,\ k=1,\ \ldots,\ n\}$$

$$=1-[P(|\xi_1|<n)]^n=1-[1-P(|\xi_1|\geqslant n)]^n$$

$$\leqslant 1-[1-nP(|\xi_1|\geqslant n)]=nP(|\xi_1|\geqslant n)\to 0$$

in view of (2). Here, we have used the inequality $(1-\alpha)^n\geqslant 1-n\alpha$ for $n\geqslant 1$, $0\leqslant\alpha\leqslant 1$.

§ 22. Central Limit Theorem

The Law of Large Numbers indicates that the share S_n/n of successes in a series of n independent trials with constant probability p of success in one trial tends (in a certain sense) to p as n increases. However, this assertion does not supply us with any information as to

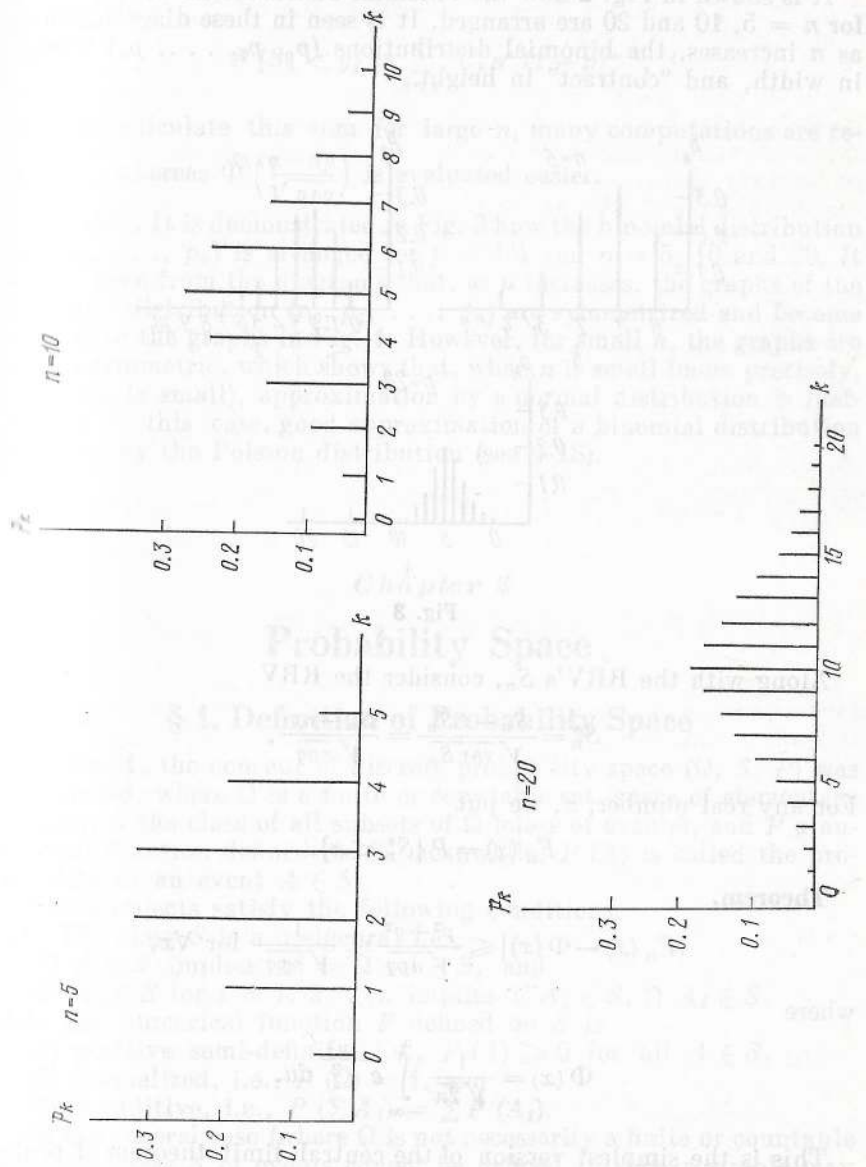

Fig. 2

how the distribution of the RRV's S_n for large n should be approximated. The answer to the last question is partially given by the so-called **Central Limit Theorems** of probability theory.

It is shown in Fig. 2 how the binomial distributions with $p = 1/2$ for $n = 5$, 10 and 20 are arranged. It is seen in these diagrams that, as n increases, the binomial distributions $(p_0, p_1, \ldots, p_n)$ "creep" in width, and "contract" in height.

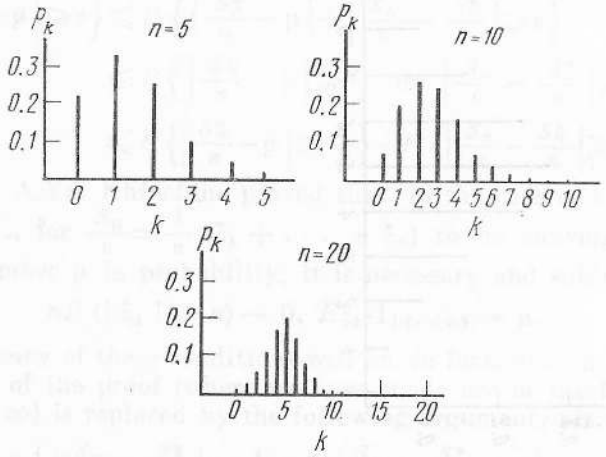

Fig. 3

Along with the RRV's S_n, consider the RRV

$$S_n^* = \frac{S_n - ES_n}{\sqrt{\operatorname{var} S_n}} = \frac{S_n - np}{\sqrt{npq}}.$$

For any real number, x, we put

$$F_n(x) = P\{S_n^* < x\}.$$

Theorem.

$$|F_n(x) - \Phi(x)| \leqslant \frac{p^2 + q^2}{\sqrt{npq}} \leqslant \frac{1}{\sqrt{npq}} \quad \text{for } \forall x,$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}}\, du.$$

This is the simplest version of the central limit theorem of probability theory. The statement of this theorem in a more general form will be given below. The function Φ is called the **normal distribution**; note that $\Phi(+\infty) = 1$.

In particular, this theorem states that, for large n,

$$P\{S_n < y\} = P\left\{S_n^* < \frac{y - np}{\sqrt{npq}}\right\} \simeq \Phi\left(\frac{y - np}{\sqrt{npq}}\right).$$

Note that

$$P\{S_n < y\} = \sum_{k<y} {}^nC_k p^k (1-p)^{n-k},$$

and, to calculate this sum for large n, many computations are required, whereas $\Phi\left(\frac{y-np}{\sqrt{npq}}\right)$ is evaluated easier.

Remark. It is demonstrated in Fig. 3 how the binomial distribution $(p_0, p_1, \ldots, p_n)$ is arranged for $p = 1/4$ and $n = 5$, 10 and 20. It can be seen from the diagrams that, as n increases, the graphs of the binomial distribution $(p_0, p_1, \ldots, p_n)$ are symmetrized and become similar to the graphs in Fig. 1. However, for small n, the graphs are more asymmetric, which shows that, when n is small (more precisely, when np is small), approximation by a normal distribution is ineffective. In this case, good approximation of a binomial distribution is given by the Poisson distribution (see § 18).

Chapter 2

Probability Space

§ 1. Definition of Probability Space

In Ch. 1, the concept of discrete probability space (Ω, S, P) was introduced, where Ω is a finite or countable set (space of elementary events), S the class of all subsets of Ω (class of events), and P a numerical function defined on S (meanwhile, $P(A)$ is called the probability of an event $A \in S$).

These objects satisfy the following conditions.

A1. The class S is a σ-algebra, i.e.,

(1) $A \in S$ implies $A^c = \Omega \setminus A \in S$, and

(2) $A_i \in S$ for $i = 1, 2, \ldots$ implies $\bigcup A_i \in S$, $\bigcap A_i \in S$.

A2. The numerical function P defined on S is

(1) positive semi-definite, i.e., $P(A) \geqslant 0$ for all $A \in S$,

(2) normalized, i.e., $P(\Omega) = 1$, and

(3) σ-additive, i.e., $P(\sum A_i) = \sum P(A_i)$.

In the general case (where Ω is not necessarily a finite or countable set), the class S of events *cannot* be defined in such a simple way (as the class of all (!) subsets of the set Ω); neither can a numerical function P be defined on this class, so that the function P is positive

semi-definite, normalized and σ-additive. It will be required that the class S of subsets of the set Ω should be only a σ-algebra.

Now, suppose that

Ω is an arbitrary set of elements ω,

S a certain nonempty class of subsets of the set Ω, and

P a numerical function defined on S.

Let Axioms A1 and A2 be fulfilled.

Probability theory yields the following interpretation of the introduced objects. The set Ω is called the space of elementary events, the elements of the class S events, and the numerical function P a probability measure, or simply probability. If A is an event, i.e., $A \in S$, then the number $P(A)$ is called the **probability** of the event A.

The triplet (Ω, S, P) is called a **probability space.**

Similarly to a discrete probability space, it follows from the listed axioms, that, in particular,

(1) $\varnothing \in S$ and $P(\varnothing) = 0$,

(2) (monotonicity of P) $A \subset B \Rightarrow P(A) \leqslant P(B)$,

(3) (strong additivity) $P(A \cup B) = P(A) + P(B) - P(AB)$,

(4) (σ-semi-additivity) $P(\bigcup A_k) \leqslant \sum P(A_k)$, and

(5) (continuity with respect to monotonic sequences) the notation $A_n \uparrow A$ means that $A_n \subset A_{n+1}$ for all n, and $A = \bigcup A_n$, whereas $A_n \downarrow A$ that $A_n \supset A_{n+1}$ for all n, and $A = \bigcap A_n$;

(a) $A_n \uparrow A \Rightarrow P(A_n) \to P(A)$,

(b) $A_n \downarrow A \Rightarrow P(A_n) \to P(A)$.

§ 2. Construction of Probability Space

In constructing a probability space (Ω, S, P), to define the function P on the whole class S of events is often a complicated task; it is more simple to define it on a certain subclass S_0 of the class S. The question arises, can the numerical function P be uniquely extended from S_0 to S?

In this connection, we introduce the following definitions (where $\mathfrak{P}(\Omega)$ denotes the class of all subsets of the set Ω).

D1. A class $S \subset \mathfrak{P}(\Omega)$ is called a **semi-algebra** of sets if

(1) $\varnothing$ and $\Omega \in S$,

(2) $A_1 \in S$, $A_2 \in S \Rightarrow A_1 \cap A_2 \in S$, and

(3) $A \in S \Rightarrow A^c = \sum_1^n A_k$ for certain pairwise disjoint subsets $A_1, \ldots, A_n$ from the class S.

D2. A class $S \subset \mathfrak{P}(\Omega)$ is called an **algebra** of sets if

(1) $\varnothing$ and $\Omega \in S$,

(2) $A \in S \Rightarrow A^c = \Omega \setminus A \in S$, and

(3) $A_1 \in S$, $A_2 \in S \Rightarrow A_1 \cup A_2 \in S$, $A_1 \cap A_2 \in S$.

Remark. This definition contains extra requirements. E.g., it is sufficient that

(1) $\Omega \in S$,

(2) $A \in S \Rightarrow A^c \in S$, and

(3) $A_1 \in S$, $A_2 \in S \Rightarrow A_1 \cap A_2 \in S$.

D3. A class $S \subset \mathfrak{P}(\Omega)$ is called a **σ-algebra** of sets if

(1) S is an algebra of sets, and

(2) $\{A_k\} \subset S \Rightarrow \bigcup A_k \in S$.

This definition of a σ-algebra is consistent with that given in A1, § 1.

It is obvious that S is a σ-algebra $\Rightarrow S$ is an algebra $\Rightarrow S$ is a semi-algebra.

D4. Let $K \subset \mathfrak{P}(\Omega)$. An algebra $S \subset \mathfrak{P}(\Omega)$ is called the **minimal algebra** containing K if $K \subset S$, and S is contained in any algebra containing K.

Such a minimal algebra exists. It is equal to the (set-theoretic) intersection of all algebras containing K. Note that the set of algebras containing K is nonempty. E.g., $\mathfrak{P}(\Omega)$ is an algeba containing K. We also note that the (set-theoretic) intersection of algebras of sets is an algebra of sets.

The minimal σ-algebra containing K is defined similarly. The proof of its existence is also similar.

D5. A numerical function P on a semi-algebra $S \subset \mathfrak{P}(\Omega)$ is called **probability** if it is

(1) positive semi-definite, i.e., $P(A) \geqslant 0$ for all $A \in S$,

(2) σ-additive, i.e., if $\{A_k\} \subset S$, $\bigcup_{k=1}^{\infty} A_k \in S$, the sets $\{A_k\}$ are pairwise disjoint, then

$$P\left(\sum A_k\right) = \sum P(A_k),$$

and

(3) normalized, i.e., $P(\Omega) = 1$.

D6. Let P_1 and P_2 be probabilities on two semi-algebras S_1 and S_2, respectively. Probability P_2 is called an **extension** of probability P_1 if

(1) $S_1 \subset S_2$,

and

(2) $P_2(A) = P_1(A)$ for all $A \in S_1$.

D7. Probability P_0 on a semi-algebra S_0 **admits a unique extension** to a semi-algebra $S \supset S_0$ if, for any pair of extensions of P_1 and P_2 from S_0 to S, $P_1(A) = P_2(A)$ for all $A \in S$.

Probability Extension Theorem. Probability P_0 on a semi-algebra S_0 admits a unique extension of P to the minimal σ-algebra S containing S_0.

This theorem (on replacing the term "probability" by the term "measure") is proved in the course of analysis.

We illustrate by some examples of semi-algebras of sets.

Example 1. Let $\Omega = [a, b]$, then the class S of sets of the form

$$(\alpha, \beta),\ (\alpha, \beta],\ [\alpha, \beta),\ [\alpha, \beta]$$

is a semi-algebra.

Example 2. The same is obtained if we put

$$\Omega = (a, b],\ S = \{(\alpha, \beta] : a \leqslant \alpha \leqslant \beta \leqslant b\}.$$

In this case,

$$A = (\alpha, \beta] \in S \Rightarrow A^c = (a, \alpha] + (\beta, b].$$

Example 3. Let $\Omega = R^1 = (-\infty, \infty)$. Then

$$S = \{(\alpha, \beta] : -\infty \leqslant \alpha \leqslant \beta < \infty\} \cup \{(\alpha, \infty) : \alpha \geqslant -\infty\}$$

is a semi-algebra, in which case $(\alpha, \beta]^c = (-\infty, \alpha] + (\beta, \infty]$ when $-\infty \leqslant \alpha \leqslant \beta < \infty$, and $(\alpha, \infty)^c = (-\infty, \alpha]$ when $\alpha \geqslant -\infty$.

Example 4. Let $a = (a_1, \ldots, a_n) \in R^n$, $b = (b_1, \ldots, b_n) \in R^n$, and $\Omega = (a, b] = (a_1, b_1] \times \ldots \times (a_n, b_n] \subset R^n$.

Then

$$S = \{(\alpha, \beta] = (\alpha_1, \beta_1] \times \ldots \times (\alpha_n, \beta_n] : a_k \leqslant \alpha_k \leqslant \beta_k \leqslant b_k\ \forall k\}$$

is a semi-algebra $\subset \mathscr{P}(\Omega)$.

Remark. Note that the minimal σ-algebra containing the semi-algebra S in Examples 1-4 is a class of Borel sets contained in Ω.

Example 5 (*direct product of semi-algebras*). Let S_k be the semi-algebra of subsets of a set Ω_k, $k = 1, \ldots, n$.

Put

$$\Omega = \Omega_1 \times \ldots \times \Omega_n,$$

$$S = S_1 \times \ldots \times S_n = \{A = A_1 \times \ldots \times A_n : A_k \in S_k\ \forall k\}.$$

Then $S \subset \mathscr{P}(\Omega)$ is a semi-algebra (equal to the direct product of the semi-algebras $S_1, \ldots, S_n$). In fact, when $n = 2$,

$$\varnothing = \varnothing \times \varnothing \in S;\quad \Omega = \Omega_1 \times \Omega_2 \in S;$$

$$A = A_1 \times A_2 \in S,$$

$$B = B_1 \times B_2 \in S \Rightarrow A \cap B = (A_1 \cap B_1) \times (A_2 \cap B_2) \in S;$$

$$A = A_1 \times A_2 \in S \Rightarrow A^c = A_1 \times A_2^c + A_1^c \times A_2 + A_1^c \times A_2^c$$

$$= \sum_j (A_1 \times A_{2j}) + \sum_i (A_{1i} \times A_2) + \sum_{ij} (A_{1i} \times A_{2j}),$$

and all the addends belong to S (where $A_1^c = \sum_i A_{1i}$, $A_{1i} \in S_1$; $A_2^c = \sum_j A_{2j}$, $A_{2j} \in S_2$).

We now give examples of probabilities on semi-algebras.

Example 6. Consider the semi-algebra S from Example 1, and put

$$P(A) = \frac{|A|}{|\Omega|}$$

for each interval $A \in S$, where $|A|$ is the length of A. Then the numerical function P on S is probability.

Example 7. Consider the semi-algebra S from Example 1 again, and also a non-decreasing function F on $(-\infty, \infty)$, so that $F(a - 0) = 0$, $F(b + 0) = 1$. Put

$$P(\alpha, \beta) = F(\beta - 0) - F(\alpha + 0)$$
$$P(\alpha, \beta] = F(\beta + 0) - F(\alpha + 0)$$
$$P[\alpha, \beta) = F(\beta - 0) - F(\alpha - 0)$$
$$P[\alpha, \beta] = F(\beta + 0) - F(\alpha - 0)$$

for $a \leqslant \alpha \leqslant \beta \leqslant b$. The numerical function P on S is probability. In the previous example, a continuous function equal to zero when $x < a$, unity when $x > b$, and linear on $[a, b]$ may be taken as F.

§ 3. Complete Probability Space

Let (Ω, S, P) be a certain probability space. A set $N \subset \Omega$ is said to be **P-zero**, or simply **zero**, if there exists a set $A \in S$ such that $N \subset A$, and $P(A) = 0$.

A probability space (Ω, S, P) is said to be **complete** if S contains all zero sets.

Consider the following procedure of completing a probability space. Let $\mathscr{N}$ be the class of zero sets. Define

(1) the class $\overline{S}$ of sets of the form $A \cup N$, where $A \in S$, $N \in \mathscr{N}$, and

(2) a numerical function $\overline{P}$ on $\overline{S}$ by putting

$$\overline{P}(A \cup N) = P(A).$$

Theorem. 1°. $\overline{S}$ is a σ-algebra coinciding with the minimal σ-algebra containing S and $\mathscr{N}$.

2°. The function $\overline{P}$ is defined correctly, being probability on $\overline{S}$.

3°. The probability space $(\Omega, \overline{S}, \overline{P})$ is complete. It is called a **completion** of the space (Ω, S, P).

Proof. 1°. It follows from the countable semi-additivity of P on S that the class $\mathscr{N}$ is closed under the countable union operation (and also countable intersection and subtraction operations on sets). Hence, the class $\overline{S}$ is closed under the countable union operation. The same class is closed under the completion operation, since

$$A \cup N \in \overline{S},\ N \subset B \in S,$$
$$P(B) = 0$$
$$\Rightarrow (A \cup N)^c = (A \cup B)^c + B \cap (A \cup N)^c = A_1 + N_1,$$

where

$$A_1 = (A \cup B)^c,\ N_1 \subset B,\ P(B) = 0 \Rightarrow (A \cup N)^c = A_1 \cup N_1,$$

with

$$A_1 \in S,\ N_1 \in \mathscr{N} \Rightarrow (A \cup N)^c \in \overline{S}.$$

Thus, S is a σ-algebra. It can be seen from the construction of $\overline{S}$ that it is the minimal σ-algebra containing S and $\mathscr{N}$.

2°. (1) The function $\overline{P}$ is defined correctly, since

$$A_1 \cup N_1 = A_2 \cup N_2 \Rightarrow A_1 \Delta A_2 = (A_1 \setminus A_2) \cup (A_2 \setminus A_1)$$
$$\subset [(A_2 \cup N_2) \setminus A_2] \cup [(A_1 \cup N_1) \setminus A_1] \subset N_1 \cup N_2$$
$$\Rightarrow P(A_1 \Delta A_2) = 0 \Rightarrow P(A_1) = P(A_2)$$
$$\Rightarrow \overline{P}(A_1 \cup N_1) = \overline{P}(A_2 \cup N_2).$$

(2) Show that $\overline{P}$ is a probability on $\overline{S}$. In fact, the function $\overline{P}$ is positive semi-definite and normalized. It remains to show that

$$\overline{P}\left(\sum_{n\geqslant 1}(A_n \cup N_n)\right) = \sum_{n\geqslant 1}\overline{P}(A_n \cup N_n).$$

Putting

$$A = \sum A_n, \quad N = \sum N_n,$$

we obtain

$$A \cup N = \sum_{n\geqslant 1}(A_n \cup N_n);$$

therefore,

$$\overline{P}\left(\sum_{n\geqslant 1}(A_n \cup N_n)\right) = \overline{P}(A \cup N) = P(A) = \sum_{n\geqslant 1} P(A_n)$$
$$= \sum_{n\geqslant 1}\overline{P}(A_n \cup N_n).$$

3°. Show that the probability space $(\Omega, \overline{S}, \overline{P})$ is complete. In fact,

$$C \subset A \cup N, \quad \overline{P}(A \cup N) = 0 \Rightarrow C \subset A \cup N; \quad P(A) = 0;$$
$$\exists B \in S: \quad N \subset B, \quad P(B) = 0 \Rightarrow C \subset A \cup B,$$
$$P(A \cup B) \leqslant P(A) + P(B) = 0 \Rightarrow C = N_1 \in \mathscr{N} \Rightarrow C = \varnothing \cup N_1 \in \overline{S}.$$

§ 4. Continuous Probability Space

Let a function $f: R^n \to R^1$ be positive semi-definite, summable, and

$$\int_{R^n} f(x)\,dx = 1.$$

For any Borel set $A \subset R^n$, we put

$$P(A) = \int_A f(x)\,dx,$$

which defines the function P on the σ-algebra of Borel sets in the space R^n. The triplet $(R^n, \mathbf{B}^n, P)$ is a probability space often called a **continuous probability space**, in which case the function f is called the **density of probability** P.

Example 1 (*classical model of probability theory, geometric definition of probability*). Let $R^n \supset \Omega$ be a Borel set of measure $|\Omega| < \infty$, and

$$f(x) = \frac{1}{|\Omega|} I_\Omega(x) = \begin{cases} \frac{1}{|\Omega|} & \text{if } x \in \Omega \\ 0 & \text{if } x \notin \Omega \end{cases}.$$

If $\Omega \supset A$ is a Borel set, then

$$P(A) = \int_A f(x)\,dx = \frac{1}{|\Omega|}\int_A dx = \frac{|A|}{|\Omega|},$$

which defines the function P on the σ-algebras of Borel sets contained in Ω. Thus, in the probability space (Ω, S, P), probability $P(A)$ is proportional to the measure of the set $A \in S$. In this case, elementary events $\omega \in \Omega$ are said to be **equally possible**, or **equally likely**.

Example 2. Let ξ be a random vector with values in R^n, and $\eta = f(\xi)$ where $f: R^n \to R^n$ is a Borel function. Let us find the density q of the distribution of the RV η in terms of the density p of the distribution of the RV ξ. We will assume that f is a smooth bijection, and $Df(x) \neq 0$, where Df is the Jacobian of the function f, i.e.,

$$Df(x) = \det\left\{\frac{\partial f_i(x)}{\partial x_j}\right\},$$

whereas

$$f(x) = \{f_1(x), \ldots, f_n(x)\}, \quad x = (x_1, \ldots, x_n).$$

We show that

$$q(y) = p(g(y)) \cdot |Dg(y)|,$$

where $g = f^{-1}$.

In fact, by the substitution formula for definite integrals,

$$P(\eta \in B) = P\{\xi \in f^{-1}(B) = A\} = \int_A p(x)\,dx$$
$$= \int_{f^{-1}(A)} p(g(y))\,|Dg(y)|\,dy = \int_B q(y)\,dy$$

for any Borel set B, which means that q is the distribution density for the RV η.

Note that, in the course of mathematical analysis, the substitution formula for definite integrals is proved for a continuous function p. However, the same also remains valid for any integrable function p if we use the Lebesgue theorem on the passage to the limit under the integral sign, and the statement that any integrable function p almost everywhere (a.e.) equals the limit of a sequence of continuous integrable functions majorized by the integrable function $|p|$.

Example 3. Let ξ and η be two random vectors with values in R^s and R^r, respectively, and the distribution of the random vector

$\zeta = (\xi, \eta)$ with values in R^{s+r} admit density $p(u, v)$, $u \in R^s$, $v \in R^r$. Since, for any Borel set $A \subset R^s$,

$$P(\xi \in A) = P(\zeta \in A \times R^r) = \int_A p(u)\, du,$$

where

$$p(u) = \int_{R^r} p(u, v)\, dv,$$

then $p(u)$ is the density distribution for the RV ξ.

Assume that the density $p(u, v)$ is continuous. Show that, for any point $u \in R^s$ such that $p(u) > 0$, there exists

$$\lim_{A_m \downarrow u} P(\eta \in B \mid \xi \in A_m) = P(\eta \in B \mid \xi = u),$$

where B is a Borel set in the space R^r, and the sequence of balls A_m of the space R^s is contractible to the point $u \in R^s$, with $P(\xi \in A_m) > 0$ for all $m = 1, 2, \ldots$. Meanwhile,

$$P(\eta \in B \mid \xi = u) = \int_B p(v \mid u)]\, dv,$$

where

$$p(v|u) = \frac{p(u, v)}{p(u)}, \text{ or } p(u, v) = p(u) \cdot p(v|u).$$

Note that

$$p(v|u) \geqslant 0, \text{ and } \int_{R^r} p(v|u)\, dv = 1,$$

due to which the function $v \to p(v \mid u)$ can be naturally defined as the density of the conditional distribution of the RV η with respect to $\xi = u$.

In fact, let $u = u_0 \in R^s$, and $p(u_0) > 0$. Since the density $p(u)$ is also continuous, we have

$$\lim_{A_m \downarrow u_0} \frac{1}{|A_m|} P(\xi \in A_m) = \lim \frac{1}{|A_m|} \int_{A_m} p(u)\, du = p(u_0),$$

$$\lim \frac{1}{|A_m|} P(\xi \in A_m, \ \eta \in B) = \lim \int_B dv \frac{1}{|A_m|} \int_{A_m} p(u, v)\, du$$

$$= \int_B p(u_0, v)\, dv,$$

where

$$|A| = \int_A du$$

denotes the volume of a ball $A \subset R^s$. Therefore, there exists

$$\lim_{A_m \downarrow u_0} P(\eta \in B | \xi \in A_m) = \int_B \frac{p(u_0, v)}{p(u_0)}\, dv.$$

§ 5. Definition of Random Variable

To the concept of random variable, the idea of a stochastic experiment is related. The latter consists in measuring a certain numerical value ξ, assuming that, for any number c, we can speak of the event "$\xi < c$" (and even of the probability of this event), due to which an RV is defined as follows.

Definition. Let (Ω, S, P) be a certain complete probability space. A numerical function $\xi : \Omega \to R^1$ such that, for any number c, the set

$$\{\xi < c\} = \{\omega : \xi(\omega) < c\} \subset \Omega$$

is an event, i.e., $\{\xi < c\} \in S$, is called a **random variable**.

The proposition below permits us to give another equivalent definition of an RV.

Lemma. The following statements are equivalent, viz.,

1°. $c \in R^1 \Rightarrow \{\xi < c\} \in S$,

and

2°. $B \in \mathbf{B} \Rightarrow \{\xi \in B\} \in S$,

where $\mathbf{B}$ is the class (σ-algebra) of Borel sets on the straight line R^1, and the notation $\{\xi \in B\}$ means the inverse image of the set B under the mapping ξ, i.e.,

$$\{\xi \in B\} = \{\omega \in \Omega : \xi(\omega) \in B\} = \xi^{-1}(B)$$

Proof. 1° follows from 2°, since

$$c \in R^1 \Rightarrow (-\infty, c) = B \in \mathbf{B} \Rightarrow \{\xi < c\} = \{\xi \in B\} \in S.$$

Let us verify that 1° $\Rightarrow$ 2°.

(1) We make use of the fact that the "inverse image" operation preserves the set-theoretic operations, which means the following. Consider a mapping $f : X \to Y$. Then

(a) $f^{-1}(\bigcup_\alpha E_\alpha) = \bigcup_\alpha f^{-1}(E_\alpha)$,

(b) $f^{-1}(\bigcap_\alpha E_\alpha) = \bigcap_\alpha f^{-1}(E_\alpha)$

and

(c) $f^{-1}(E^c) = (f^{-1}(E))^c$,

where $f^{-1}(E)$ is the inverse image of a set $E \subset Y$. Now, if $\mathscr{E}$ is a certain class of subsets of the set Y, then it follows from (a)-(c) that

$$\sigma f^{-1}(\mathscr{E}) = f^{-1}(\sigma \mathscr{E}),$$

$\sigma\mathscr{E}$ denoting the minimal σ-algebra containing the class $\mathscr{E}$, and

$$f^{-1}(\mathscr{E}) = \{f^{-1}(E) : E \in \mathscr{E}\}.$$

In particular, if $\mathscr{E}$ is the σ-algebra of subsets of the set Y (and then $\sigma\mathscr{E} = \mathscr{E}$), $f^{-1}(\mathscr{E})$ is the σ-algebra of subsets of the set X.

(2) Putting $f = \xi$, and assuming that $\mathscr{E}$ is the class of half-lines $(-\infty, c)$, we obtain

$$B \in \mathbf{B} \Rightarrow \xi^{-1}(B) \in \xi^{-1}(\mathbf{B}) = \xi^{-1}(\sigma\mathscr{E}) = \sigma\xi^{-1}(\mathscr{E}) \subset S,$$

since

$$\sigma\mathscr{E} = \mathbf{B}, \quad \text{and} \quad \xi^{-1}(\mathscr{E}) \subset S.$$

The proved lemma means that a mapping $\xi : \Omega \to R^1$ is an RV if and only if the inverse image of any Borel set is an event. The following statement yields a sufficient (and necessary) condition for a mapping $\xi : \Omega \to R^1$ to be an RV.

Theorem. Let $\mathscr{E} \subset \mathscr{P}(R^1)$, and $\sigma(\mathscr{E}) = \mathbf{B}$. For a mapping $\xi : \Omega \to R^1$ to be an RV, it is necessary and sufficient that

$$\{\xi \in E\} \in S$$

for all $E \subset \mathscr{E}$.

Proof.

(1) ξ is an RV, $E \in \mathscr{E} \Rightarrow \xi$ is an RV, $E \in \mathbf{B} \Rightarrow \{\xi \in E\} \in S$ according to the lemma.

(2) $B \in \mathbf{B} \Rightarrow \{\xi \in B\} = \xi^{-1}(B) \in \xi^{-1}(\mathbf{B}) = \xi^{-1}(\sigma\mathscr{E}) = \sigma\xi^{-1}(\mathscr{E}) \subset S$, since $\xi^{-1}(\mathscr{E}) = \{\xi^{-1}(E) : E \in \mathscr{E}\} \subset S$.

§ 6. Operations over RV's

A numerical function given on the straight line is said to be Borel if the inverse image of each Borel set is a Borel set.

1°. Let ξ be an RV, and φ a Borel function. Then $\eta = \varphi(\xi)$ is an RV (or, briefly, a Borel function of an RV is an RV; in particular, a continuous function of an RV is an RV). E.g., if ξ is an RV, then

$$\xi^2, \quad |\xi|, \quad \xi^+ = \max(0, \xi)$$

is also an RV. If ξ is an RV, and $\xi(\omega) \neq 0$ for all ω, then $1/\xi$ is also an RV, since $1/\xi = \varphi(\xi)$, where the function $\varphi : R^1 \to R^1$ defined as

$$\varphi(x) = \begin{cases} 1/x & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

is Borel.

A function $\varphi: R^n \to R^1$ is said to be Borel (of n variables) if the inverse image of each Borel set in R^1 is a Borel set in R^n.

2°. Let $\varphi: R^n \to R^1$ be a Borel function, and $\xi_1, \ldots, \xi_n$ RV's, then $\eta = \varphi(\xi_1, \ldots, \xi_n)$ is an RV. In particular, the sum, difference, product and quotient of two RV's (provided the denominator does not vanish) is again an RV.

3°. If $\{\xi_n, n \geqslant 1\}$ is a sequence of RV's, then $\sup \xi_n$, $\inf \xi_n$, $\overline{\lim}\, \xi_n$ and $\underline{\lim}\, \xi_n$ are also RV's. In particular, if, for each $\omega \in \Omega$, there exists $\lim \xi_n(\omega) = \xi(\omega)$, then ξ is an RV.

Proof.

1°. $B \in \mathbf{B} \Rightarrow (\eta \in B) = (\xi \in \varphi^{-1}(B)\} \in S$, since $\varphi^{-1}(B) \in \mathbf{B}$, and ξ is an RV.

2°. Let $\mathbf{B}^n$ be the class (σ-algebra) of Borel sets in R^n, and $\mathscr{E}$ a subclass of the class $\mathbf{B}^n$, consisting of sets of the form $B_1 \times \ldots \times B_n \subset R^n$, where $B_k \in \mathbf{B} \forall k$. Then $\sigma(\mathscr{E}) = \mathbf{B}^n$. We have

$$B \in \mathbf{B} \Rightarrow \{\eta \in \mathbf{B}\} = \{(\xi_1, \ldots, \xi_n) \in \varphi^{-1}(B)\} = \xi^{-1}(A) \in \xi^{-1}(\mathbf{B}^n)$$
$$= \xi^{-1}(\sigma\mathscr{E}) = \sigma\xi^{-1}(\mathscr{E}) \subset S,$$

where $\xi = (\xi_1, \ldots, \xi_n)$, $A = \varphi^{-1}(B) \in \mathbf{B}^n$. The latter inclusion follows from $\xi^{-1}(\mathscr{E}) \subset S$. In fact, let $E = B_1 \times \ldots \times B_n \in \mathscr{E}$. Then

$$\xi^{-1}(E) = \{\xi \in E\} = \{\xi_1 \in B_1\} \cap \ldots \cap \{\xi_n \in B_n\} \in S.$$

3°. The statement follows from the relations

$$\{\sup \xi_n \leqslant c\} = \bigcap \{\xi_n \leqslant c\},$$

$$\inf \xi_n = -\sup(-\xi_n), \qquad \overline{\lim}\, \xi_n = \inf_n \sup_p \xi_{n+p},$$

and

$$\underline{\lim}\, \xi_n = \sup_n \inf_p \xi_{n+p}.$$

§ 7. Equivalence and Almost Certain Convergence of RV's

Definition 1. Two RV's ξ and η are said to be equivalent, which is denoted by $\xi \sim \eta$, if $P\{\xi \neq \eta\} = 0$.

A certain property depending on elementary events $\omega \in \Omega$ is said to be fulfilled **almost certainly** (a.c.) if it is fulfilled for all ω except those forming a set (event) of zero probability. In particular, two RV's are said to be equivalent if they coincide a.c.

Lemma. Let a function $\xi : \Omega \to R^1$ coincide with an RV η a.c. Then ξ is an RV.

Proof. Since the sets

$$\{\omega : \xi(\omega) < c\} \quad \text{and} \quad \{\omega : \eta(\omega) < c\}$$

may differ only by a set of zero probability, then the measurability of the second (due to the completeness of the probability space) entails that of the first.

Definition 2. A sequence $\{\xi_n\}$ of RV's is said to **converge a.c.** to an RV ξ if $\lim \xi_n(\omega) = \xi(\omega)$ a.c. This fact is denoted by $\xi_n \overset{\text{a.c.}}{\to} \xi$.

Note that $\xi_n \overset{\text{a.c.}}{\to} \xi$ is equivalent to

$$P\{\lim \xi_n = \xi\} = 1;$$

therefore, in this case, we speak of convergence with probability one, and write

$$P \lim \xi_n = \xi.$$

Theorem 1. Let $\xi_1, \xi_2, \ldots$ be n RV's, and $\lim \xi_n(\omega) = \xi(\omega)$ a.c. Then ξ is an RV.

Proof. Let Ω_0 be the set of elementary events ω for which $\xi_n(\omega) \to \xi(\omega)$, $N = \Omega \setminus \Omega_0$. Then $P(N) = 0$. Put $\xi'_n(\omega) = \xi_n(\omega)$ on Ω_0 and $\xi'_n(\omega) = 0$ on N. It follows from the lemma that ξ'_n is an RV. Similarly, we define ξ' in terms of ξ. Then $\lim \xi'_n = \xi'$, and, by 3°, § 6, ξ' is an RV, whereas, according to the lemma, ξ is an RV.

Theorem 2. Let for each $n = 1, 2, \ldots$, RV's ξ_n and η_n be equivalent. Then the following pairs of RV's viz.,

$$\sup \xi_n \text{ and } \sup \eta_n, \quad \inf \xi_n \text{ and } \inf \eta_n,$$

$$\overline{\lim}\, \xi_n \text{ and } \overline{\lim}\, \eta_n, \quad \underline{\lim}\, \xi_n \text{ and } \underline{\lim}\, \eta_n$$

are equivalent.

Proof. Put $N_k = \{\omega : \xi_k(\omega) \neq \eta_k(\omega)\}$. Then $\sup \xi_n(\omega) \neq \sup \eta_n(\omega)$ on a set $N \subset \bigcup_{k \geqslant 1} N_k$. Now, the equivalence of the pair of RV's $\sup \xi_n$ and $\sup \eta_n$ follows from

$$P(\bigcup N_k) \leqslant \sum P(N_k) = 0,$$

and also from the completeness of the probability space.

The equivalence of the remaining pairs of RV's is shown similarly.

Theorem 3. Let, for each $k = 1, \ldots, n$, RV's ξ_k and η_k be equivalent, and $\varphi: R^n \to R^1$ a Borel function. Then the RV's

$$\xi = \varphi(\xi_1, \ldots, \xi_n) \text{ and } \eta = \varphi(\eta_1, \ldots, \eta_n)$$

are equivalent.

Proof. Put $N_k = \{\omega : \xi_k(\omega) \neq \eta_k(\omega)\}$. Then $\xi(\omega) \neq \eta(\omega)$ on a set $N \subset N_1 \cup \ldots \cup N_n$, but $P(N_1 \cup \ldots \cup N_n) \leqslant \sum_1^n P(N_k) = 0$, and, due to the completeness of the probability space, $P(N) = 0$. Here, we have used the fact that ξ and η are RV's (see § 6).

§ 8. Independence

A. Independence of classes of events.

*D1. Two classes of events $\mathscr{E}_1$ and $\mathscr{E}_2$ are said to be **independent** if

$$P(A_1 A_2) = P(A_1) \cdot P(A_2)$$

for all $A_1 \in \mathscr{E}_1$ and $A_2 \in \mathscr{E}_2$. In particular, if the classes $\mathscr{E}_1$ and $\mathscr{E}_2$ are independent, and each of them consists of one event A_1 and A_2, respectively, then the events A_1 and A_2 are said to be independent.

D2. Let $\{\mathscr{E}_i : i \in I\}$ be a family of classes of events. Classes of this family are said to be independent if, for any finite set $J \subset I$,

$$P(\bigcap_{j \in J} A_j) = \prod_{j \in J} P(A_j)$$

for all $A_j \in \mathscr{E}_j$, $j \in J$.

The following theorem yields a sufficient condition of independence of σ-algebras of events. A class of events $\mathscr{E}$ is called a **multiplicative**, or π-class, if it is closed with respect to the finite intersection operation, i.e., $A \in \mathscr{E}$, $B \in \mathscr{E}$ implies $AB \in \mathscr{E}$.

Theorem 1. Let $\{\mathscr{E}_i : i \in I\}$ be a family of classes of events. Assume that each class of this family is multiplicative. Then, for σ-algebras $\{\sigma(\mathscr{E}_i) : i \in I\}$ to be independent, the independence of classes $\{\mathscr{E}_i : i \in I\}$ is sufficient (and necessary).

Corollary. Let each of the independent classes $\mathscr{E}_i$ consist of one event A_i. It is clear that the class $\mathscr{E}_i$ is multiplicative, and

$$\sigma(\mathscr{E}_i) = \{\Omega, \varnothing, A_i, A_i^c\}.$$

Therefore, the independence of the events $A_1, \ldots, A_n$ implies that of the events $A_1^{\sigma_1}, \ldots, A_n^{\sigma_n}$ for $\sigma_i = 0$ or 1, where

$$A^{\sigma} = \begin{cases} A & \text{if } \sigma = 0 \\ A^c & \text{if } \sigma = 1 \end{cases}.$$

Proof of Theorem 1. It suffices to consider the case where the family $\{\mathscr{E}_i : i \in I\}$ is finite. Fix $i \in I$, and denote by S the class of events A_i such that

$$P(\bigcap_{j \in I} A_j) = \prod_{j \in I} P(A_j)$$

for all $A_j \in \mathscr{E}_j$, $i \neq j$. It is sufficient to show that $S \supset \sigma(\mathscr{E}_i)$. It is clear that $S \supset \mathscr{E}_i$, while

(1) $\Omega \in S$,

(2) $S \ni B_1 \subset B_2 \in S$ implies $B_2 \setminus B_1 \in S$, and

(3) $\{B_n : n = 1, 2, \ldots\} \subset S$, $B_i \cap B_j = \varnothing$ when $i \neq j$ implies $\bigcup_{n \geqslant 1} B_n \in S$.

Note that a class $S \subset \mathscr{P}(\Omega)$ satisfying conditions (1)-(3) is sometimes called a σ-additive, or λ-class. Thus, a λ-class S contains the π-class $\mathscr{E}_i$. It remains to make use of the following statement.

Lemma 1. If a λ-class S_1 contains a π-class S_0, then $S_1 \supset \sigma(S_0)$.

Proof. Denote the minimal λ-class containing S_0 by S^* (note that the set-theoretic intersection of λ-classes is a λ-class), and put

$$S(B) = \{A : A \in S^*, AB \in S^*\} \text{ for } B \in S^*.$$

* D stands for *definition*.

It is easy to verify that $S(B)$ is a λ-class for all $B \in S^*$. If, however, $B \in S_0$, then $S_0 \subset S(B) \subset S^*$, and, since S^* is the minimal λ-class containing S_0, then $S(B) = S^*$, which means that $A \in S^*$, $B \in S_0$ implies $AB \in S^*$. Hence,

$$S = \{A : A \in S^*,\ AB \in S^* \text{ for all } B \in S^*\} \supset S_0.$$

Besides, S is a λ-class as the intersection of λ-classes $S(B)$ in $B \in S^*$. Thus $S_0 \subset S \subset S^*$, S is a λ-class, and S^* the minimal λ-class containing S_0, whence $S = S^*$. But this means that $A \in S^*$, $B \in S^*$ implies $AB \in S^*$. Therefore, S^* is a π-class and λ-class simultaneously, i.e., S^* is a σ-algebra. Taking into account the definition of the class S^*, we have $S^* = \sigma(S_0)$, and, finally, $S_1 \supset S^* = \sigma(S_0)$.

B. Independence of RV's. Two RRV's ξ_1 and ξ_2 are said to be **independent** if, for any two numbers c_1 and c_2, the events

$$(\xi_1 < c_1) = \{\omega : \xi_1(\omega) < c_1\} \text{ and } (\xi_2 < c_2) = \{\omega : \xi_2(\omega) < c_2\}$$

are independent. Now, let $\{\xi_i : i \in I\}$ be a family of RV's given on one and the same probability space. RV's of this family are said to be independent (mutually) if, for any set of numbers $\{c_i : i \in I\}$, the events

$$\{(\xi_i < c_i) : i \in I\}$$

are independent. Recall that this means the independence of the events

$$\{(\xi_j < c_j) : j \in J\}$$

for any finite set $J \subset I$. Two random vectors $\xi = (\xi_1, \ldots, \xi_m)$ and $\eta = (\eta_1, \ldots, \eta_n)$ are independent if, for any collection of numbers $c = (c_1, \ldots, c_m)$ and $d = (d_1, \ldots, d_n)$, the events

$$(\xi < c) = (\xi_1 < c_1, \ldots, \xi_m < c_m) \text{ and}$$

$$(\eta < d) = (\eta_1 < d_1, \ldots, \eta_n < d_n)$$

are independent. The independence of the random vectors of the family $\{\xi_i : i \in I\}$ is defined similarly.

Theorem 2. Let

$$\varphi_k : R_k^n \to R_k^m$$

be a Borel function, and ξ_k a random vector of dimension n_k, $k = 1, \ldots, s$. If the random vectors $\xi_1, \ldots, \xi_s$ are independent, then the random vectors

$$\eta_k = \varphi_k(\xi_k), \quad k = 1, \ldots, s,$$

are also independent.

In particular, if the RV's $\xi_1, \ldots, \xi_n$ are independent, then the RV's

$$\eta_1 = \varphi_1(\xi_{i_1}, \ldots, \xi_{j_1}),$$
$$\eta_2 = \varphi_2(\xi_{i_2}, \ldots, \xi_{j_2}),$$
$$\ldots\ldots\ldots\ldots$$
$$\eta_s = \varphi_s(\xi_{i_s}, \ldots, \xi_{j_s})$$

are also independent when $1 \leqslant i_1 \leqslant j_1 \leqslant i_2 \leqslant j_2 < \ldots < i_s \leqslant j_s \leqslant n$.

Proof of Theorem 2 is based on the following statement permitting to give an equivalent difinition of RV independence.

Lemma 2. Let $\{\xi_i : \Omega \to R^{n_i},\ i \in I\}$ be a family of random vectors. Denote the class of Borel sets in the space R^n by $\mathbf{B}^n$. Then the independence of random vectors $\{\xi_i : i \in I\}$ is equivalent to that of the events

$$\{(\xi_i \in B_i) : i \in I\}$$

for all $B_i \in \mathbf{B}^{n_i}$, $i \in I$.

Proof of Lemma 2. It suffices to show that the independence of the random vectors $\{\xi_i : i \in I\}$ implies that of the indicated events. Put

$$\mathscr{E}_i = \{(\xi_i < c) : c \in R^{n_i}\}.$$

$\mathscr{E}_i$ is, obviously, a multiplicative class, and $\sigma(\mathscr{E}_i) = \mathbf{B}^{n_i}$. The statement of Lemma 2 now follows from Theorem 1.

Proof of Theorem 2. The independence of events

$$\{\eta_k \in B_k\}, \quad k = 1, \ldots, s,$$

for $B_k \in \mathbf{B}^{m_k}$ follows from the equality

$$\{\eta_k \in B_k\} = \{\xi_k \in \varphi_k^{-1}(B_k)\},$$

independence of the random vectors $\xi_1, \ldots, \xi_s$, Lemma 2 and $\varphi_k^{-1}(B_k) = A_k \in \mathbf{B}^{n_k} \forall k$.

§ 9. Mathematical Expectation, Distribution Function, Distribution Density

D1. An RV is said to be **discrete** if it assumes a finite or countable number of values.

D2. Let a discrete RV ξ assume values from $\{x_n\}$, and $A_n = \{\xi = x_n\}$. The discrete RV ξ is said to be **integrable** (or **summable**) if the series

$$\sum x_n P(A_n)$$

converges absolutely. The sum of the series is denoted by

$$E\xi = \sum_n x_n P(A_n),$$

and called the **mathematical expectation** (ME) of the integrable and discrete RV ξ.

D3. An RV ξ is said to be **integrable** if there exists a sequence $\{\xi_n\}$ of integrable and discrete RV's, which is uniformly convergent to the RV ξ (as a function of $\omega \in \Omega$). Then the number

$$(*) \qquad E\xi = \lim_{n\to\infty} E\xi_n$$

is called the mathematical expectation of the integrable RV ξ.

Proposition 1. Definition D1 is correct, i.e.,

(1) the limit (*) exists,

(2) does not depend on the choice of a sequence $\{\xi_n\}$, and

(3) for a discrete RV, the definitions of integrability and ME coincide with D2.

The latter proposition, as well as those below, are proved in the course of analysis when studying the Lebesgue integral.

The ME of an RV ξ is also denoted by

$$E\xi = \int_\Omega \xi(\omega)\, P(d\omega) = \int_\Omega \xi\, dP,$$

which is the Lebesgue integral of the summable function $\xi : \Omega \to R^1$ with respect to measure P.

If L is a collection of integrable RV's, and the functional $E : L \to R^1$ the ME operation, then

(1) E is linear,

(2) E is monotonic, i.e., $\xi_1 \leqslant \xi_2$ (which means that $\xi_1(\omega) \leqslant \xi_2(\omega)\ \forall\omega$) implies $E\xi_1 \leqslant E\xi_2$,

(3) $\xi_1 \in L$, $\xi_2 \in L$, $P(\xi_1 = \xi_2) = 1$ implies $E\xi_1 = E\xi_2$, and

(4) if two RV's ξ_1 and ξ_2 are integrable and independent, then the RV $\xi_1 \cdot \xi_2$ is integrable, while

$$E(\xi_1 \cdot \xi_2) = E\xi_1 \cdot E\xi_2$$

(for discrete RV's, the proof is the same as in § 11, Ch. 1; in the general case, the statement follows from D3).

D4. If ξ is an RV, then the function F defined by the expression

$$F(x) = P(\xi < x) \quad \forall x \in R^1$$

is called the **distribution function** (DF) of the RV ξ. To stress the dependence of the DF F on the RV ξ, it is often denoted by F_ξ.

Note that DF F

(1) is nondecreasing on R^1,

(2) continuous on the left, and

(3) $F(-\infty) = 0$, $F(+\infty) = 1$.

Any function F satisfying conditions (1)-(3) is the DF of a certain RV on a certain probability space (see Example 7, § 2).

Proposition 2. Let ξ be an RV, F_ξ its DF, and φ a Borel function. Then

$$E\varphi(\xi) = \int_{-\infty}^{\infty} \varphi(x)\, dF_\xi(x),$$

while $\varphi(\xi) \in L$ if and only if there exists the integral on the right-hand side. In particular, if $\varphi(x) = x$, then

$$E\xi = \int_{-\infty}^{\infty} x\, dF_\xi(x),$$

and both sides exist simultaneously.

D5. Let F be a DF, and there exist a summable, positive semi-definite function $f : R^1 \to R^1$ such that

$$F(x) = \int_{-\infty}^{x} f(u)\, du \quad \forall x \in R^1.$$

Such a function f is called the **density** of the DF F.

In particular, if there exists the density f of the DF F of an RV ξ, then

$$E\xi = \int_{-\infty}^{\infty} x f(x)\, dx,$$

and both sides exist simultaneously.

D6. If $\xi_1, \ldots, \xi_n$ are n RV's, then the function F defined by the relation

$$F(x) = F(x_1, \ldots, x_n) = P(\xi_1 < x_1, \ldots, \xi_n < x_n)\ \forall x$$
$$= (x_1, \ldots, x_n) \in R^n$$

is called the **joint distribution function** of the RV's $\xi_1, \ldots, \xi_n$, or the DF of the random vector $\xi = (\xi_1, \ldots, \xi_n)$. It is obvious that the RV's $\xi_1, \ldots, \xi_n$ are independent if and only if

$$F(x_1, \ldots, x_n) = F_1(x_1) \ldots F_n(x_n)\ \forall x = (x_1, \ldots, x_n) \in R^n,$$

where

$$F_k(x_k) = P(\xi_k < x_k)\ \forall k.$$

Let there exist a summable, positive semi-definite function $f : R^n \to R^1$ such that

$$F(x) = \int_{-\infty}^{x} f(u)\, du \quad \forall x \in R^n,$$

i.e., $\forall (x_1, \ldots, x_n) \in R^n$,

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f(u_1, \ldots, u_n)\, du_1 \ldots du_n.$$

Such a function f is called the **density** of the DF F. It is obvious that if the DF's $F_1, \ldots, F_n$ of RV's $\xi_1, \ldots, \xi_n$ admit densities $f_1, \ldots, f_n$, respectively, and the RV's $\xi_1, \ldots, \xi_n$ are independent, then the joint DF E of the RV's $\xi_1, \ldots, \xi_n$ admits density f, while

$$f(x_1, \ldots, x_n) = f_1(x_1) \ldots f_n(x_n)\ \forall (x_1, \ldots, x_n) \in R^n.$$

Remark. Let us list the basic properties of the space L of integrable RV's, and also of the mathematical expectation operation.

1°. **The Beppo Levi Theorem.** Let a sequence $\{\xi_n\} \subset L$ satisfy the condition

$$\xi_n \uparrow \xi \text{ a.c. and } E\xi_n \leqslant c = \text{const.}$$

Then $\xi \in L$, and $E\xi = \lim\limits_{n\to\infty} E\xi_n$.

2°. **The Lebesgue Theorem.** Let $\{\xi_n\} \subset L$ again, and

$$\xi_n \to \xi \text{ a.c., } |\xi_n| \leqslant \xi_0 \in L \text{ a.c.}$$

Then $\xi \in L$ and $E\xi = \lim\limits_{n\to\infty} E\xi_n$.

3°. **The Fatou Lemma.** Let $\{\xi_n\} \subset L$,

$$\xi_n \to \xi \text{ a.c., and } E|\xi_n| \leqslant c = \text{const.}$$

Then $\xi \in L$, and $E|\xi| \leqslant c$.

4°. **The Completeness Theorem.** The space of integrable RV's with those coinciding a.c. regarded as equivalent, is complete with respect to the norm $\|\xi\| = E|\xi|$.

§ 10. Characteristic Functions

Let ξ be an RV. Then $e^{it\xi}$ also is an RV (t is a real number). Meanwhile, by definition, $\xi + i\eta$ is a complex RV if ξ and η are real RV's. Since the modulus of the RV $e^{it\xi}$ is bounded (by unity), it is integrable, i.e., there exists $Ee^{it\xi}$. The function

$$\varphi(t) = Ee^{it\xi}$$

of a real variable t is called the **characteristic function** of the RV ξ. Note that

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx}\, dF(x),$$

where

$$F(x) = F_\xi(x) = P(\xi < x).$$

We now indicate the basic simple properties of the CF.

Property 1. If, for a certain integer $n \geqslant 1$, there exists $E\xi^n$ (i.e., $E|\xi|^n < \infty$), then the CF $\varphi(t)$ is differentiable n times, and

$$\varphi^{(n)}(0) = i^n E\xi^n.$$

This property permits us to determine the moments of an RV from its CF.

Property 2. Let ξ_1 and ξ_2 be two independent RV's, φ_1 and φ_2 the corresponding CF's, and $\xi = \xi_1 + \xi_2$. Then

$$\varphi(t) = Ee^{it\xi} = \varphi_1(t)\cdot\varphi_2(t),$$

i.e., the CF of the sum of independent RV's, equals the product of the CF's of the RV's to be added.

Property 3 (Uniqueness Theorem). The CF

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx}\, dF(x), \quad -\infty < t < \infty,$$

uniquely determines the DF F.

Proofs.

Property 1.

A. Note first of all that $E|\xi|^n < \infty$ implies $E|\xi|^k < \infty$ for all $0 \leqslant k \leqslant n$. Show that there exists the derivative $\varphi'(t)$, and $\varphi'(0) = iE\xi$, for which we consider the expression

$$\frac{\varphi(t+h)-\varphi(t)}{h} = \int_{-\infty}^{\infty} e^{itx}\,\frac{e^{ihx}-1}{h}\, dF(x), \quad h \neq 0,$$

and make use of the inequality

$$\left|\frac{e^{ihx}-1}{h}\right| \leqslant |x|$$

obtained from

$$|e^{it}-1| = \left|i\int_0^t e^{ix}\, dx\right| \leqslant |t|, \quad -\infty < t < \infty.$$

Since the function

$$\alpha_h(x) = e^{itx}\,\frac{e^{ihx}-1}{h}$$

satisfies the inequality

$$|\alpha_h(x)| \leqslant |x|,$$

and $\int\limits_{-\infty}^{\infty} |x|\, dF(x) = E|\xi| < \infty$, it follows from the Lebesgue theorem that there exists

$$\lim_{h\to 0} \int_{-\infty}^{\infty} \alpha_h(x)\, dF(x)$$

equal to

$$i\int_{-\infty}^{\infty} e^{itx} x\, dF(x).$$

Therefore, there exists the derivative

$$\varphi'(t) = i\int_{-\infty}^{\infty} xe^{itx}\, dF(x).$$

In particular,

$$\varphi'(0) = i\int_{-\infty}^{\infty} x\,dF(x) = iE\xi.$$

B. Let us verify by induction that, for each integer k, $1 \leqslant k \leqslant n$, the CF is differentiable k times, and

$$\varphi^{(k)}(t) = i^k \int_{-\infty}^{\infty} x^k e^{itx}\,dF(x).$$

For $k = 1$, this formula has already been proved. Let it be valid for a certain $k < n$. Consider the expression

$$\frac{\varphi^{(k)}(t+h) - \varphi^{(k)}(t)}{h} = i^k \int_{-\infty}^{\infty} x^k e^{itx}\,\frac{e^{ihx}-1}{h}\,dF(x),$$

and make use of the inequality

$$|\alpha_h(x)| \leqslant |x|^{k+1}$$

which is valid for the function

$$\alpha_h(x) = i^k x^k e^{itx}\,\frac{e^{ihx}-1}{h}, \quad h \neq 0.$$

Since

$$\int_{-\infty}^{\infty} |x|^{k+1}\,dF(x) = E\,|\xi|^{k+1} < \infty,$$

by the Lebesgue theorem again, there exists

$$\lim_{h\to 0} \int_{-\infty}^{\infty} \alpha_h(x)\,dF(x)$$

equal to

$$i^{k+1} \int_{-\infty}^{\infty} x^k e^{itx} \cdot x\,dF(x).$$

Therefore, there exists the derivative

$$\varphi^{(k+1)}(t) = i^{k+1} \int_{-\infty}^{\infty} x^{k+1} e^{itx}\,dF(x).$$

Property 2 follows from the representation

$$\varphi(t) = E\,e^{it\xi_1} \cdot e^{it\xi_2},$$

independence of the RV's $e^{it\xi_1}$, $e^{it\xi_2}$, and multiplicativity of the ME operation E.

Property 3. Let the CF's coincide for two DF's F and G, i.e.,

$$(1) \qquad \int_{-\infty}^{\infty} e^{itx}\,dF(x) = \int_{-\infty}^{\infty} e^{itx}\,dG(x), \quad -\infty < t < \infty.$$

To show that F and G coincide at the points of continuity, we shall make use of the following two statements.

(i) Any function continuous on $[-l, l]$, $l > 0$, and assuming the same values at the ends of this interval is the limit of a uniformly convergent series of trigonometric polynomials of the form

$$(2) \qquad T(x) = \sum_k c_k \exp\left[i\pi x\,\frac{k}{l}\right]$$

(with a finite number of terms).

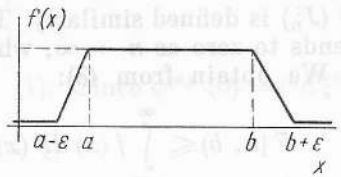

Fig. 4

(ii) It follows from (i) that, for any trigonometric polynomial of form (2),

$$\int_{-\infty}^{\infty} T(x)\,dF(x) = \int_{-\infty}^{\infty} T(x)\,dG(x).$$

Proceeding from these statements, we shall now show that

$$(3) \qquad \int_{-\infty}^{\infty} f(x)\,dF(x) = \int_{-\infty}^{\infty} f(x)\,dG(x)$$

for any function f of the form indicated in Fig. 4, viz., for f continuous on $(-\infty, \infty)$, equal to unity on $[a, b]$, zero outside $[a-\varepsilon, b+\varepsilon]$, and linear on $[a-\varepsilon, a]$ and $[b, b+\varepsilon]$, where $a \leqslant b$, $\varepsilon > 0$.

To this end, we consider a sequence of numbers $0 < \varepsilon_n \leqslant 1$, $\varepsilon_n \to 0$, and sequence of intervals $J_n = [-n, n]$, beginning with the one containing $[a-\varepsilon, b+\varepsilon]$. For each of such numbers n, we select, by statement (i), a trigonometric polynomial $T_n(x)$, so that

$$|f(x) - T_n(x)| \leqslant \varepsilon_n \leqslant 1 \quad \text{for} \quad x \in J_n.$$

Since the function $T_n(x)$ considered on $(-\infty, \infty)$ is periodic (with period $|J_n| = 2n$), and $|f(x)| \leqslant 1$,

$$|T_n(x)| \leqslant 2 \text{ for } -\infty < x < \infty.$$

Now, making use of statement (ii), we obtain

$$\left|\int_{-\infty}^{\infty} f(x)\,dF(x) - \int_{-\infty}^{\infty} f(x)\,dG(x)\right| = \left|\int_{-n}^{n} f(x)\,dF(x) - \int_{-n}^{n} f(x)\,dG(x)\right|$$

$$\leqslant \left|\int_{-n}^{n} T_n(x)\,dF(x) - \int_{-n}^{n} T_n(x)\,dG(x)\right| + 2\varepsilon_n$$

$$\leqslant \left| \int_{-\infty}^{\infty} T_n(x)\, dF(x) - \int_{-\infty}^{\infty} T_n(x)\, dG(x) \right| + 2F(J_n^c) + 2G(J_n^c) + 2\varepsilon_n$$
$$= 2[F(J_n^c) + G(J_n^c) + \varepsilon_n],$$

where

$$F(J_n) = F(n+0) - F(-n), \quad J_n^c = (-\infty, \infty) \setminus J_n,$$
$$F(J_n^c) = 1 - F(J_n) = F(-n) + 1 - F(n+0).$$

$G(J_n^c)$ is defined similarly. The right-hand side of these inequalities tends to zero as $n \to \infty$, which proves (3).

We obtain from (3):

$$F[a, b) \leqslant \int_{-\infty}^{\infty} f(x)\, dF(x) = \int_{-\infty}^{\infty} f(x)\, dG(x) \leqslant G[a-\varepsilon, b+\varepsilon),$$

which implies that $F[a, b) \leqslant G[a, b)$ for all points a and b of continuity of the functions F and G. Similarly, $G[a, b) \leqslant F[a, b)$, i.e.,

$$F[a, b) = G[a, b),$$

or

$$F(b) - F(a) = G(b) - G(a).$$

Hence, $F(x) = G(x)$ for all x.

§ 11. One-Dimensional Normal Distribution

(a) Recall that

$$\int_{-\infty}^{\infty} e^{-u^2}\, du = \sqrt{\pi}.$$

Therefore, the function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}}\, du, \quad \Phi(+\infty) = 1, \tag{1}$$

is a DF.

(b) Find the CF of this distribution, viz.,

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx}\, d\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx - \frac{x^2}{2}}\, dx$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}t^2 - \frac{1}{2}(x-it)^2\right\} dx$$
$$= e^{-\frac{t^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} \int_{D} e^{-\frac{u^2}{2}}\, du,$$

where we have put $u = x - it$, and integrated with respect to the straight line

$$D = \{u : u = x - it, \quad -\infty < x < \infty\}$$

parallel to the real axis. Shifting the path of integration onto the real axis, leaving the value of the latter integral unaltered, and making use of the equality $\Phi(+\infty) = 1$, we obtain, finally, that

$$\boxed{\varphi(t) = e^{-\frac{t^2}{2}}} \tag{2}$$

(c) Let ξ be an RV with distribution (1). Since $\varphi^{(k)}(0) = i^k E\xi^k$, we find from (2):

$$E\xi = 0, \quad \operatorname{var} \xi = 1.$$

For the RV $\eta = \sigma\xi + \mu$, $\sigma > 0$, we have

$$E\eta = \mu, \quad \operatorname{var} \eta = \sigma^2,$$
$$F_\eta(x) = P\{\eta < x\} = P\left\{\xi < \frac{x-\mu}{\sigma}\right\} = \Phi\left(\frac{x-\mu}{\sigma}\right)$$
$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}}\, du,$$
$$\varphi_\eta(t) = Ee^{it\eta} = e^{it\mu} Ee^{i(t\sigma)\xi} = \exp\left\{it\mu - \frac{1}{2}\sigma^2 t^2\right\}.$$

(d) The distribution

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{(u-\mu)^2}{2\sigma^2}\right\} du, \quad -\infty < x < \infty,$$

is called the **normal distribution**, and denoted by $N(\mu, \sigma^2)$. The density of such a distribution equals

$$f(x) = F'(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < \infty.$$

The normal distribution $N(0, 1)$ is said to be **standard.**

(e) Let two RV's ξ_1 and ξ_2 be independent, and each have the normal distribution $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. Then the RV $\xi = \xi_1 + \xi_2$ has the normal distribution $N(\mu, \sigma^2)$, where $\mu = \mu_1 + \mu_2$, $\sigma^2 = \sigma_1^2 + \sigma_2^2$, which follows from the CF of the RV ξ being equal to

$$\varphi(t) = Ee^{it\xi} = Ee^{it\xi_1} \cdot Ee^{it\xi_2} = \exp\left\{it\mu_1 - \frac{1}{2}\sigma_1^2 t^2\right\}$$
$$\times \exp\left\{it\mu_2 - \frac{1}{2}\sigma_2^2 t^2\right\} = \exp\left\{it\mu - \frac{1}{2}\sigma^2 t^2\right\}$$

because of (c), and since, by Property 3, § 10, a CF uniquely determines the DF.

(f) Let an RV ξ have the normal distribution $N(\mu, \sigma^2)$. Then the RV $\eta = a\xi + b$ has the distribution $N(a\mu + b, a^2\sigma^2)$. In fact, the CF of the RV η equals

$$\varphi_\eta(t) = Ee^{it\eta} = e^{itb}Ee^{i(at)\xi} = \exp\left\{it(a\mu + b) - \frac{1}{2}a^2\sigma^2t^2\right\}.$$

§ 12. Gamma Distribution

(a) Recall that, by definition, the gamma function is

$$\Gamma(\lambda) = \int_0^\infty u^{\lambda-1}e^{-u}\,du, \quad \lambda > 0.$$

In particular,

$$\Gamma(n+1) = n!, \quad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}, \quad \Gamma(\lambda+1) = \lambda\Gamma(\lambda).$$

The distribution

$$F(x) = \frac{1}{\Gamma(\lambda)}\int_0^x u^{\lambda-1}e^{-u}\,du, \quad x \geqslant 0, \quad F(+\infty) = 1$$

is called the **gamma distribution**, and denoted by $G(\lambda)$. If λ is an integer, then this distribution is sometimes called the **Erlang distribution.**

(b) Find the CF of the gamma distribution $G(\lambda)$, viz.,

$$\varphi(t) = \int_{-\infty}^\infty e^{itx}\,dF(x) = \frac{1}{\Gamma(\lambda)}\int_0^\infty e^{itx-x}\cdot x^{\lambda-1}\,dx$$

$$= (1-it)^{-\lambda}\cdot\frac{1}{\Gamma(\lambda)}\int_R u^{\lambda-1}e^{-u}\,du,$$

where $u = (1 - it)\,x$, and integration is performed with respect to the ray

$$R = \{u : u = (1 - it)\,x,\ x \geqslant 0\}$$

placed in the right-hand half-plane. Carrying the path of integration into the positive part of the real axis, leaving the value of the latter integral unaltered, and taking into account the equality $F(+\infty) = 1$, we obtain, finally, that

$$\boxed{\varphi(t) = (1-it)^{-\lambda}}.$$

(c) Let ξ be an RV with the gamma distribution $G(\lambda)$. Since $\varphi^{(k)}(0) = i^kE\xi^k$, and $\operatorname{var}\xi = E\xi^2 - (E\xi)^2$,

$$E\xi = \operatorname{var}\xi = \lambda.$$

(d) Let two RV's ξ_1 and ξ_2 be independent, and each have the gamma distribution $G(\lambda_1)$ and $G(\lambda_2)$, respectively. Then the RV $\xi = \xi_1 + \xi_2$ has the gamma distribution $G(\lambda)$, where $\lambda = \lambda_1 + \lambda_2$, which follows from the CF of the RV ξ being equal to

$$\varphi(t) = Ee^{it\xi_1}\cdot Ee^{it\xi_2} = (1-it)^{-\lambda_1}\cdot(1-it)^{-\lambda_2} = (1-it)^{-\lambda},$$

because of (b), and since a CF uniquely determines the DF.

§ 13. Beta and Other Distributions

A. (a) Recall that the beta function is defined by the formula

$$B(\lambda_1, \lambda_2) = \int_0^1 u^{\lambda_1-1}(1-u)^{\lambda_2-1}\,du, \quad \lambda_1 > 0,\ \lambda_2 > 0.$$

The relation between the beta and gamma functions is determined by the equality

$$B(\lambda_1, \lambda_2) = \frac{\Gamma(\lambda_1)\cdot\Gamma(\lambda_2)}{\Gamma(\lambda_1+\lambda_2)}.$$

A distribution of an RV ξ on the interval $[0, 1]$, given by the equality

$$F(x) = \frac{1}{B(\lambda_1, \lambda_2)}\int_0^x u^{\lambda_1-1}(1-u)^{\lambda_2-1}\,du, \quad 0 \leqslant x \leqslant 1,$$

is called the **beta distribution**, and denoted by $\mathrm{Be}(\lambda_1, \lambda_2)$, while

$$E\xi = \frac{\lambda_1}{\lambda_1+\lambda_2}, \quad \operatorname{var}\xi = \frac{\lambda_1\lambda_2}{(\lambda_1+\lambda_2)^2(\lambda_1+\lambda_2+1)}.$$

In fact,

$$E\xi = \frac{1}{B(\lambda_1, \lambda_2)}\int_0^1 x^{\lambda_1}(1-x)^{\lambda_2-1}\,dx = \frac{B(\lambda_1+1, \lambda_2)}{B(\lambda_1, \lambda_2)}$$

$$= \frac{\Gamma(\lambda_1+1)\cdot\Gamma(\lambda_2)\cdot\Gamma(\lambda_1+\lambda_2)}{\Gamma(\lambda_1+\lambda_2+1)\cdot\Gamma(\lambda_1)\cdot\Gamma(\lambda_2)} = \frac{\lambda_1}{\lambda_1+\lambda_2},$$

where we have made use of the formula $\Gamma(\lambda + 1) = \lambda\Gamma(\lambda)$, $\lambda > 0$.

Thus,

$$E\xi^2 = \frac{1}{B(\lambda_1, \lambda_2)}\int_0^1 x^{\lambda_1+1}(1-x)^{\lambda_2-1}\,dx = \frac{B(\lambda_1+2, \lambda_2)}{B(\lambda_1, \lambda_2)}$$

$$= \frac{(\lambda_1+1)\lambda_1}{(\lambda_1+\lambda_2+1)(\lambda_1+\lambda_2)},$$

from which $\operatorname{var}\xi = E\xi^2 - (E\xi)^2$ is determined.

(b) If ξ_1 and ξ_2 are two independent RV's with the gamma distributions $G(\lambda_1)$ and $G(\lambda_2)$, respectively, then the RV

$$\xi = \frac{\xi_1}{\xi_1 + \xi_2}$$

has the beta distribution Be (λ_1, λ_2).

In fact, for $0 < x < 1$, we have

$$F(x) = P\{\xi < x\} = P\left\{\frac{1}{\xi_1 + \xi_2} < x\right\}$$

$$= \frac{1}{\Gamma(\lambda_1)\cdot\Gamma(\lambda_2)} \iint\limits_{\frac{u}{u+v}<x} u^{\lambda_1 - 1} e^{-u} \cdot v^{\lambda_2 - 1} e^{-v}\, du\, dv, \quad u > 0,\ v > 0.$$

Substitute

$$\frac{u}{u+v} = w, \quad u + v = y, \text{ or } u = wy, \quad v = y - wy.$$

Then the Jacobian is

$$\frac{\partial(u, v)}{\partial(w, y)} = \begin{vmatrix} \frac{\partial u}{\partial w} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial w} & \frac{\partial v}{\partial y} \end{vmatrix} = \begin{vmatrix} y & w \\ -y & 1-w \end{vmatrix} = (1-w)\, y + yw = y.$$

Therefore,

$$F(x) = \frac{1}{\Gamma(\lambda_1)\,\Gamma(\lambda_2)} \iint\limits_{\substack{0 \leqslant w < x \\ y \geqslant 0}} (wy)^{\lambda_1 - 1} [y(1-w)]^{\lambda_2 - 1} e^{-y} y\, dy\, dw$$

$$= \frac{1}{\Gamma(\lambda_1)\,\Gamma(\lambda_2)} \int\limits_{0 \leqslant w < x} w^{\lambda_1 - 1} (1-w)^{\lambda_2 - 1}\, dw \int\limits_{y \geqslant 0} y^{\lambda_1 + \lambda_2 - 1} e^{-y}\, dy$$

$$= \frac{\Gamma(\lambda_1 + \lambda_2)}{\Gamma(\lambda_1)\,\Gamma(\lambda_2)} \int\limits_0^x w^{\lambda_1 - 1} (1-w)^{\lambda_2 - 1}\, dw.$$

B. Other distributions of continuous type.

(1) The **Chi-square distribution (Pearson distribution).** This is the distribution of the RV

$$\chi^2 = \xi_1^2 + \ldots + \xi_n^2,$$

where $\xi_1, \ldots, \xi_n$ are independent, and each has the normal distribution $N(0, 1)$. This distribution is denoted by $P(n)$, or $\chi^2(n)$. The number n is called the **number of degrees of freedom.** The distribution of the RV χ^2 coincides with that of the RV 2ξ, where ξ has the gamma distribution $G(n/2)$, while

$$E\chi^2 = n, \text{ var } \chi^2 = 2n.$$

(2) The **Student distribution** (t-distribution). Let two RV's ξ and η be independent, and have the distributions $N(0, 1)$ and $P(n)$,

respectively. Then the RV distribution

$$t = \frac{\xi}{(\eta/n)^{1/2}}$$

is called the Student distribution (or t-distribution), and denoted by $S(n)$. The density of this distribution is of the form

$$c_n \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}},$$

where the normalizing factor c_n equals

$$c_n = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \cdot \frac{1}{\sqrt{n\pi}},$$

$$Et = 0, \text{ var } t = \frac{n}{n-2} \text{ when } n > 2.$$

When $n = 1$, the Student distribution is called the **Cauchy distribution.** Note that the RV $\left(1 + \frac{t^2}{n}\right)^{-1}$ has the beta distribution Be $\left(\frac{n}{2}, \frac{1}{2}\right)$.

(3) The **Snedecor-Fisher distribution.** This is the RV distribution

$$\eta = \frac{\xi_1}{k_1} \div \frac{\xi_2}{k_2},$$

where the RV's ξ_1 and ξ_2 are independent, and have distributions $P(k_1)$ and $P(k_2)$, respectively. It is denoted by F_{k_1, k_2}. The density of this distribution is of the form

$$c(k_1, k_2)\, x^{\frac{k_1}{2} - 1} \left(1 + \frac{k_1}{k_2} x\right)^{-\frac{k_1 + k_2}{2}},$$

where $c(k_1, k_2)$ is a normalizing factor equal to

$$\frac{\Gamma\left(\frac{k_1 + k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right)\cdot\Gamma\left(\frac{k_2}{2}\right)} \left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}},$$

and

$$E\eta = \frac{k_2}{k_2 - 2}, \quad \text{var } \eta = \frac{2k_2^2 (k_1 + k_2 - 2)}{k_1 (k_2 - 2)(k_2 - 4)}.$$

Note that if ξ has the distribution Be $\left(\frac{k_1}{2}, \frac{k_2}{2}\right)$, then the RV

$$\eta = \frac{k_2 \xi}{k_1 (1 - \xi)}$$

has the distribution F_{k_1, k_2}.

An RV with the Snedecor-Fisher distribution is sometimes denoted by F, and the Fisher distribution spoken of. The symbol F was introduced by Snedecor in honour of R. A. Fisher. R. A. Fisher himself used the value $\frac{1}{2}\ln F$.

(4) The **logarithmic normal distribution** corresponds to the distribution of an RV ξ for which $\ln \xi$ has the normal distribution $N(0, 1)$.

(5) The **exponential distribution** is of the form

$$F(x) = P(\xi < x) = 1 - e^{-ax}, \quad x \geqslant 0, \ a > 0.$$

It is clear that the RV $a\xi$ has the gamma distribution $G(1)$.

(6) The **Weibull distribution**

$$P(\xi < x) = 1 - e^{-ax^{\lambda}}, \ x \geqslant 0, \ a > 0, \ \lambda > 0.$$

It is called the **Rayleigh distribution** in the case $\lambda = 2$.

(7) The **uniform distribution** on the interval $[0, 1]$

$$P(\xi < x) = x, \quad 0 \leqslant x \leqslant 1.$$

It is clear that ξ has the beta distribution $\mathrm{Be}(1, 1)$.

(8) Quite often, a **three-parameter family** of distributions is used. It is determined by the family of densities

$$f(x \mid \alpha, \beta, \gamma) = c(\alpha, b, \gamma)\, x^{\alpha} e^{-bx^{\gamma}},$$

where $c(\alpha, \beta, \gamma)$ is a normalizing factor.

§ 14. Uniform Distribution

A. Consider the uniform distribution on $(0, 1)$, i.e., distribution with density

$$f(x) = \begin{cases} 1, & x \in (0, 1) \\ 0, & x \notin (0, 1) \end{cases}.$$

The DF is continuous, equal to zero on $(-\infty, 0]$, unity on $[1, \infty)$, and is linear on $(0, 1)$. If an RV ξ is uniformly distributed on $(0, 1)$, then

$$E\xi = \frac{1}{2}, \quad E\xi^2 = \frac{1}{3}, \quad \operatorname{var}\xi = E\xi^2 - (E\xi)^2 = \frac{1}{12}.$$

$$\varphi_{\xi}(t) = Ee^{it\xi} = \int_0^1 e^{itx}\, dx = \frac{e^{it} - 1}{it}.$$

B. Assume that n RV's $\xi_1, \ldots, \xi_n$ are independent, add each has the uniform distribution on $(0, 1)$. Put

$$U = \min_i \xi_i, \ V = \max_i \xi_i,$$

$$D = V - U, \ R = U/V.$$

Find the distribution of these RV's and $\operatorname{cor}(U, V)$.

C. For the RV V, we have

$$P\{V < x\} = P\{\bigcap_1^n (\xi_i < x)\} = \prod_1^n P(\xi_i < x) = x^n$$

if $0 < x < 1$. Thus, the RV V taking values from $(0, 1)$ has distribution density

$$\boxed{f_V(x) = nx^{n-1}}, \quad 0 < x < 1,$$

$$EV = \int_0^1 x \cdot nx^{n-1}\, dx = \frac{n}{n+1} = 1 - \frac{1}{n+1},$$

$$EV^2 = \int_0^1 x^2 \cdot nx^{n-1}\, dx = \frac{n}{n+2} = 1 - \frac{2}{n+2},$$

$$\operatorname{var} V = EV^2 - (EV)^2 = \frac{n}{(n+2)(n+1)^2}.$$

D. For the RV U, we have

$$P\{U < x\} = 1 - P\{U \geqslant x\} = 1 - P\{\bigcap_1^n (\xi_i \geqslant x)\}$$

$$= 1 - \prod_1^n P(\xi_i \geqslant x) = 1 - (1-x)^n$$

if $0 < x < 1$. Thus, the RV U taking values from $(0, 1)$ has distribution density

$$\boxed{f_U(x) = n(1-x)^{n-1}}, \quad 0 < x < 1.$$

Note that the distributions of U and $1 - V$ coincide (which can be otherwise obtained from $1 - V = \min_i (1 - \xi_i)$, where the RV's $1 - \xi_1, \ldots, 1 - \xi_n$ are independent and uniformly distributed on $(0, 1)$), with

$$EU = 1 - EV = \frac{1}{n+1}, \quad \operatorname{var} U = \operatorname{var} V.$$

E. Before finding the distribution of the RV's D and R, we find the joint distribution of the RV's U and V. The vector-valued random variable (U, V) takes values from

$$\Delta = \{(u, v) : 0 < u \leqslant v < 1\}.$$

If $(u, v) \in \Delta$, then

$$P\{U < u, \ V < v\} = P\{V < v\} - P\{U \geqslant u, \ V < v\}$$

$$= v^n - P\{\bigcap_i^n (u \leqslant \xi_i < v)\}$$

$$= v^n - \prod_1^n P(u \leqslant \xi_i < v) = v^n - (v-u)^n.$$

Thus, the RV (U, V) assuming values in Δ has density

$$\boxed{\frac{\partial^2}{\partial u\,\partial v} P\{U<u,\ U<v\} = f(u, v) = n(n-1)(v-u)^{n-2}},$$

where $0<u\leqslant v<1$.

F. For the RV $D = V - U$, we have

$$P\{D<x\} = 1 - P\{D\geqslant x\} = 1 - \iint\limits_{\substack{x<v-u<1\\ u>0}} f(u, v)\,du\,dv$$

$$= 1 - n(n-1)\int_x^1 dv \int_0^{v-x} (v-u)^{n-2}\,du$$

$$= 1 - n\int_x^1 (v^{n-1} - x^{n-1})\,dv = nx^{n-1} - (n-1)x^n,$$

if $0<x<1$. Thus, the RV D assuming values from $(0, 1)$ has distribution density

$$f_D(x) = n(n-1)[x^{n-2} - x^{n-1}].$$

It is assumed that $n\geqslant 2$ (otherwise $D=0$). Thus,

$$ED = EV - EU = \frac{n-1}{n+1},$$

$$ED^2 = n(n-1)\int_0^1 (x^{n-2} - x^{n-1})x^2\,dx$$

$$= n(n-1)\left[\frac{1}{n+1} - \frac{1}{n+2}\right] = \frac{n(n-1)}{(n+1)(n+2)},$$

$$\operatorname{var} D = ED^2 - (ED)^2 = \frac{2(n-1)}{(n+1)^2(n+2)}.$$

G. For the RV $R = U/V$, we have

$$P\{R<x\} = \iint\limits_{u/v<x} f(u, v)\,du\,dv = \int_0^1 dv \int_0^{vx} n(n-1)(v-u)^{n-2}\,du$$

$$= n\int_0^1 [v^{n-1} - (v - vx)^{n-1}]\,dv = 1 - (1-x)^{n-1}$$

if $0<x<1$. Thus, the RV R assuming values from $(0, 1)$ has distribution density

$$\boxed{f_R(x) = (n-1)(1-x)^{n-2}}, \quad 0<x<1.$$

It is assumed that $n\geqslant 2$ (otherwise $R = 1$ when $n = 1$).

Note that, according to Item D, the RV R has the same distribution as the RV $\min(\xi_1, \ldots, \xi_{n-1})$. Therefore,

$$ER = \frac{1}{n}, \quad \operatorname{var} R = \frac{n-1}{n^2(n+1)}.$$

H. To determine $\operatorname{cor}(U, V)$, we can make use of the formulas

$$\operatorname{var} U = \operatorname{var} V, \quad D = V - U,$$

$$\operatorname{var} D = \operatorname{var} V + \operatorname{var} U - 2\operatorname{cor}(U, V)$$

$$= 2\operatorname{var} V\,[1 - \operatorname{cor}(U, V)],$$

$$\frac{\operatorname{var} D}{2\operatorname{var} V} = \frac{n-1}{n} = 1 - \frac{1}{n}.$$

Finally, we obtain

$$\boxed{\operatorname{cor}(U, V) = \frac{1}{n}}.$$

§ 15. VRV Characteristic Function

Let $\xi = (\xi_1, \ldots, \xi_r)$ be a VRV. Then $e^{i(t,\xi)}$ is a complex-valued RV, where $t = (t_1, \ldots, t_r)$, and $(t, \xi) = \sum_1^r t_k\xi_k$. Since the modulus of $e^{i(t,\xi)}$ is bounded (by unity), it is integrable. The function

$$\varphi(t) = Ee^{i(t,\xi)}, \quad t\in R^r,$$

defined on R^r is called the **characteristic function of the VRV** ξ. Note that

$$\varphi(t) = \int_{R^r} e^{i(t,x)}dF(x),$$

where

$$F(x) = F_\xi(x) = P\{\xi<x\} = P\{\xi_1<x_1, \ldots, \xi_r<x_r\}$$

for $x = (x_1, \ldots, x_r)\in R^r$.

We now indicate the basic simple properties of the CF of a VRV.

Property 1. Put

$$\xi^n = \xi_1^{n_1}\ldots\xi_r^{n_r},$$

where the components of the vector $n = (n_1, \ldots, n_r)$ are non-negative integers. Similarly, if $D_k = \partial/\partial t_k$ is the partial differentiation operator over functions of $t = (t_1, \ldots, t_r)\in R^r$, then, by definition,

$$D = (D_1, \ldots, D_r), \quad D^n = D_1^{n_1}\ldots D_r^{n_r} = \left(\frac{\partial^{|n|}}{\partial t_1^{n_1}\ldots\partial t_1^{n_r}}\right),$$

$$|n| = n_1 + \ldots + n_r.$$

If there exists $E\xi^n$ (i.e., $E\,|\,\xi^n\,| < \infty$), then $D^n\varphi\,(t)$ also exists, and

$$D^n\varphi\,(0) = i^{|n|}\cdot E\xi^n.$$

This property permits us to determine the product moments of a VRV from the CF.

Property 2. Let ξ_1 and ξ_2 be two independent VRV's of the same dimension, φ_1 and φ_2 the corresponding CF's, and $\xi = \xi_1 + \xi_2$. Then

$$\varphi\,(t) = Ee^{i(t,\xi)} = \varphi_1\,(t)\cdot\varphi_2\,(t),$$

i.e., the CF of the sum of independent VRV's equals the product of the CF's of the RV's to be added.

Property 3 (Uniqueness Theorem). The CF

$$\varphi\,(t) = \int\limits_{R^r} e^{i(t,x)}\,dF\,(x),\quad t \in R^r,$$

determines the DF F uniquely.

Proofs of these properties are similar to those in the one-dimensional case (see § 10).

§ 16. Multidimensional Normal Distribution

A. Consider a VRV $\xi = (\xi_1, \ldots, \xi_r)$ such that the RRV's $\xi_1, \ldots, \xi_r$ are i.i.d. according to $N\,(0, 1)$. For two elements

$$x = (x_1, \ldots, x_r) \quad \text{and} \quad y = (y_1, \ldots, y_r)$$

from R^r, the inequality $x < y$ is, by definition, equivalent to the system of inequalities

$$x_1 < y_1, \ldots, x_r < y_r.$$

Then the DF of the VRV ξ equals

$$\Phi\,(x) = P\,\{\xi < x\} = P\,\{\bigcap_1^r (\xi_k < x_k)\} = \prod_1^r P\,(\xi_k < x_k)$$

$$= \prod_1^r \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x_k} e_k^{-u_k^2/2}\,du_k$$

$$= (2\pi)^{-\frac{r}{2}} \int\limits_{u_1<x_1} \cdots \int\limits_{u_r<x_r} e^{-\frac{1}{2}(u_1^2+\ldots+u_r^2)}\,du_1 \ldots du_r$$

$$= (2\pi)^{-r/2} \int\limits_{u<x} e^{-\frac{1}{2}(u,\,u)}\,du.$$

The distribution density equals

$$f\,(x) = \frac{\partial^r}{\partial x_1 \ldots \partial x_r}\,\Phi\,(x) = (2\pi)^{-\frac{r}{2}} \exp\left\{-\frac{1}{2}(x,\,x)\right\},\quad x \in R^r,$$

whereas the CF of the RV ξ is

$$\varphi_\xi\,(t) = Ee^{i(t,\,\xi)} = Ee^{it_1\xi_1} \ldots Ee^{it_r\xi_r}$$

$$= e^{-\frac{t_1^2}{2}} \ldots e^{-\frac{t_r^2}{2}} = \exp\left\{-\frac{1}{2}(t,\,t)\right\}.$$

B. Let σ be an arbitrary real non-singular matrix of order $r \times r$, and μ a certain fixed vector from R^r. We will represent vectors from R^r as column vectors. Find the distribution of the VRV

$$\eta = \sigma\xi + \mu.$$

Note *a priori* one general statement. If the VRV ξ has distribution density $f\,(x)$, then the VRV η has distribution density $g\,(x)$ which is determined by the equality

$$g\,(x) = |\det\sigma^{-1}|\cdot f\,[\sigma^{-1}\,(x-\mu)].$$

The idea of the proof of this statement consists in the following. If $S_1, S_2, \ldots$ is a sequence of balls with centres at a point x, their radii tending to zero, then, by denoting by $v\,(M)$ the volume (measure) of a measurable set $M \subset R^r$, we have

$$g\,(x) = \lim \frac{P\,\{\eta \in S_n\}}{v\,(S_n)} = \lim \frac{P\,\{\xi \in \sigma^{-1}\,(S_n-\mu)\}}{v\,(\sigma^{-1}\,(S_n-\mu))}\,\frac{v\,(\sigma^{-1}\,(S_n-\mu))}{v\,(S_n)},$$

where

$$\sigma^{-1}\,(S_n-\mu) = \{\sigma^{-1}\,(z-\mu) \in R^r : z \in S_n\}.$$

Since $\sigma^{-1}\,(S_n-\mu)$ is an ellipse with volume

$$v\,(\sigma^{-1}\,(S_n-\mu)) = v\,(\sigma^{-1}S_n) = |\det\sigma^{-1}|\cdot v\,(S_n),$$

the statement follows from the sequence of ellipses $\sigma^{-1}\,(S_n-\mu)$ being contractible to the point $\sigma^{-1}\,(x-\mu)$; therefore,

$$\lim \frac{P\,\{\xi \in \sigma^{-1}\,(S_n-\mu)\}}{v\,(\sigma^{-1}\,(S_n-\mu))} = f\,(\sigma^{-1}\,(x-\mu)).$$

Note that the indicated assertion can be extended (and proved similarly) to the case of any one-to-one mapping with Jacobian different from zero.

In our case,

$$f\,(x) = (2\pi)^{-\frac{r}{2}} \exp\left\{-\frac{1}{2}(x,\,x)\right\};$$

therefore,

$$g(x) = (2\pi)^{-\frac{r}{2}} |\det \sigma|^{-1} \exp\left\{-\frac{1}{2}(\sigma^{-1}(x-\mu), \sigma^{-1}(x-\mu))\right\}$$
$$= \frac{1}{(2\pi)^{\frac{r}{2}} |A|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(A^{-1}(x-\mu), x-\mu)\right\},$$

where $A = \sigma\sigma'$ is a positive definite matrix (σ' being the transpose of σ), $|A| = \det A$.

A distribution with such density $g(x)$ on R^r is called a **multidimensional normal distribution**, and denoted by $N_r(\mu, A)$.

Note that if, instead of σ, we took a matrix $\sigma_1 = \sigma H$, where H is an orthogonal matrix, then the distribution for the random vector

$$\eta_1 = \sigma_1 \xi + \mu$$

would not change, since

$$A_1 = \sigma_1 \sigma_1' = \sigma H H' \sigma' = \sigma\sigma' = A.$$

Geometrically, this means that if we performed a transformation of rotation or symmetry type over a random vector with the multidimensional normal distribution $N_r(0, 1)$, then the obtained random vector would have the same normal distribution.

C. Find the CF of the random vector η. We have

$$\varphi_\eta(t) = Ee^{i(t, \eta)} = E\exp\{i(t, \sigma\xi) + i(t, \mu)\}$$
$$= e^{i(t, \mu)} Ee^{i(\sigma' t, \xi)} = e^{i(t, \mu)} \varphi_\xi(\sigma' t),$$

and, finally,

$$\varphi_\eta(t) = \exp\left\{i(t, \mu) - \frac{1}{2}(\sigma' t, \sigma' t)\right\} = \exp\left\{i(t, \mu) - \frac{1}{2}(At, t)\right\}.$$

D. Let us clarify the meaning of parameters μ and A determining the normal distribution $N_r(\mu, A)$ of the RV η. Since $E\xi = 0$, where

$$E\xi = (E\xi_1, \ldots, E\xi_r)',$$

we have

$$E\eta = \mu.$$

Representing vectors from R^r as column vectors again, and making use of the transposing sign $'$, we obtain from $\eta = \sigma\xi + \mu$:

$$(\eta - \mu)(\eta - \mu)' = \sigma\xi \cdot (\sigma\xi)' = \sigma(\xi\xi')\sigma',$$

and, since $E\xi\xi' = I$, where I is the unit matrix, we get, finally,

$$\mu = E\eta,$$
$$A = \sigma\sigma' = E(\eta - \mu)(\eta - \mu)'.$$

The vector μ is called the **mean vector**. Since the matrix A is representable in the form

$$A = \{\mathrm{cov}(\eta_i, \eta_j)\},$$

where

$$\mathrm{cov}(\eta_i, \eta_j) = E(\eta_i - \mu_i)(\eta_j - \mu_j),$$

it is called a **covariance matrix.**

E. Assume that two VRV's ξ_1 and ξ_2 are independent, and have distributions $N_r(\mu_1, A_1)$ and $N_r(\mu_2, A_2)$, respectively. Then the VRV $\xi = \xi_1 + \xi_2$ has a normal distribution $N_r(\mu, A)$, where

$$\mu = \mu_1 + \mu_2, \quad A = A_1 + A_2.$$

In fact, according to Item B, the CF of the RV ξ equals

$$\varphi_\xi(t) = \varphi_{\xi_1}(t) \cdot \varphi_{\xi_2}(t) = \exp\left\{i(t, \mu_1) - \frac{1}{2}(A_1 t, t)\right\}$$
$$\times \exp\left\{i(t, \mu_2) - \frac{1}{2}(A_2 t, t)\right\} = \exp\left\{i(t, \mu) - \frac{1}{2}(At, t)\right\}.$$

It now remains to make use of the fact that a CF determines the distribution uniquely (see Property 3, § 15).

F. Let an RV ξ have distribution $N_r(\mu, A)$, and C be a nonsingular real matrix of order $r \times r$. Then the RV $\eta = C\xi$ has a distribution $N_r(C\mu, CAC')$.

In fact, the CF of the RV η equals

$$\varphi_\eta(t) = Ee^{i(t, C\xi)} = Ee^{i(C't, \xi)} = \varphi_\xi(C't)$$
$$= \exp\left\{i(C't, \mu) - \frac{1}{2}(AC't, C't)\right\}$$
$$= \exp\left\{i(t, C\mu) - \frac{1}{2}(CAC't, t)\right\}.$$

It now remains to make use of the fact that a CF determines the distribution uniquely.

G. Let an RV ξ have a distribution $N_r(\mu, A)$, and c be a column vector from R^r. Then the RV $\eta = (c, \xi)$ has the distribution $N_1((c, \mu), (Ac, c))$.

In fact, the CF of the RRV η equals

$$\varphi_\eta(t) = Ee^{it\eta} = Ee^{it(c, \xi)} = Ee^{i(tc, \xi)} = \varphi_\xi(tc)$$
$$= \exp\left\{i(tc, \mu) - \frac{1}{2}(Atc, tc)\right\}$$
$$= \exp\left\{it(c, \mu) - \frac{1}{2}t^2(Ac, c)\right\}.$$

H. The following statement generalizes the previous two, and is proved similarly. Let an RV ξ have a distribution $N_r(\mu, A)$, and C be a real rectangular matrix of order $k \times r$ and rank $k \leqslant r$. Then the RV $\eta = C\xi$ has the distribution $N_k(C\mu, CAC')$.

§ 17. Markov Inequality

$$P\{|\xi| \geqslant \varepsilon\} \leqslant \frac{E|\xi|^k}{\varepsilon^k}$$

for all RV's ξ and numbers $k \geqslant 0$, $\varepsilon > 0$.

Proof is the same as for the discrete case (see Ch. 1).

Corollary. Putting $k = 2$, and replacing ξ by $\xi - E\xi$, we obtain the Chebyshev inequality

$$P\{|\xi - E\xi| \geqslant \varepsilon\} \leqslant \frac{\operatorname{var}\xi}{\varepsilon^2}.$$

§ 18. Jensen Inequality

A. A numerical function φ defined on the straight line R^1 is convex if

$$\varphi(\alpha x + \beta y) \leqslant \alpha\varphi(x) + \beta\varphi(y)$$

for all numbers x, y, α and β, $\alpha \geqslant 0$, $\beta \geqslant 0$, $\alpha + \beta = 1$, which is equivalent to the statement that, for each point $x_0 \in R^1$, there exists $c = c(x_0) \in R^1$ such that

$$\varphi(x) - \varphi(x_0) \geqslant c(x - x_0) \tag{1}$$

for all $x \in R^1$ (see Fig. 5).

Let ξ be an RRV, and φ a function which is convex on R^1. Then the following **Jensen inequality** is valid, viz.,

$$\boxed{\varphi(E\xi) \leqslant E\varphi(\xi)}$$

assuming that the ME's involved exist. The proof is obtained from (1) if we put $x = \xi$, $x_0 = E\xi$, and apply the ME operation E to both sides of the inequality.

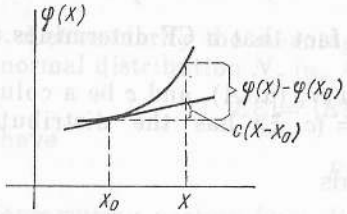

Fig. 5

B. A set $M \subset R^n$ is said to be **convex** if, together with any two of its points, it also contains the line segment which joins them. A numerical function φ defined on a convex set $M \subset R^n$ is said to be convex if

$$\varphi(\alpha x + \beta y) \leqslant \alpha\varphi(x) + \beta\varphi(y)$$

for all x and y from M, and all numbers α and β such that $\alpha \geqslant 0$, $\beta \geqslant 0$, $\alpha + \beta = 1$.

This requirement is equivalent to the fact that, for any point $x_0 \in M$, there exists $c = c(x_0) \in R^n$ such that

$$\varphi(x) - \varphi(x_0) \geqslant (c,\ x - x_0) \tag{2}$$

for all $x \in M$, where the right-hand side is the scalar product of the vectors c and $x - x_0$.

Now, let ξ be a VRV with values in a convex set $M \subset R^n$, and φ a numerical convex function defined on M. If there exist $E\xi$ and $E\varphi(\xi)$, then the Jensen inequality holds, viz.,

$$\varphi(E\xi) \leqslant E\varphi(\xi).$$

Proof is obtained from (2) if we put $x = \xi$, $x_0 = E\xi$, and apply the operation E to both sides of the inequality.

Remark. Assume that a numerical function φ defined on a convex set $M \subset R^n$ is strictly convex, i.e.,

$$\varphi(\alpha x + \beta y) < \alpha\varphi(x) + \beta\varphi(y)$$

for all x and $y \neq x$ from M, and all numbers α and β such that $\alpha > 0$, $\beta > 0$, $\alpha + \beta = 1$. This requirement is equivalent to the fact that inequality (2) is valid strictly for all $x \neq x_0$. If $P(\xi \neq E\xi) > 0$, then $\varphi(E\xi) < E\varphi(\xi)$. In fact, putting $x_0 = E\xi$, and $\eta = \varphi(\xi) - \varphi(x_0) - (c,\ \xi - x_0)$, we obtain $\varphi(E\xi) = E\varphi(\xi) \Leftrightarrow E\eta = 0$, $\eta \geqslant 0$ a.c. $\Leftrightarrow \eta = 0$ a.c. $\Leftrightarrow \xi = E\xi$ a.c. $\Leftrightarrow P(\xi \neq E\xi) = 0$.

§ 19. Kolmogorov Inequality

Assume that RRV's $\xi_1, \ldots, \xi_n$ are independent, and $E\xi_k = 0$, $E\xi_k^2 < \infty$ for all k. Then

$$\boxed{P\{\max_{1 \leqslant k \leqslant n} |\xi_1 + \ldots + \xi_k| \geqslant a\} \leqslant \frac{\sum_1^n E\xi_k^2}{a^2}}$$

for all $a > 0$.

Proof. Put

$$s_0 = 0, \quad s_k = \xi_1 + \ldots + \xi_k, \quad A = \{\max_{1 \leqslant k \leqslant n} |s_k| \geqslant a\},$$

and

$$A_k = \{\max_{0 \leqslant i < k} |s_i| < a,\ |s_k| \geqslant a\}.$$

Since $A = \sum_1^n A_k$, we have

$$\sum_1^n E\xi_k^2 = Es_n^2 \geqslant Es_n^2 I_A = \sum_1^n Es_n^2 I_{A_k} = \sum_1^n E[s_k + (s_n - s_k)]^2 I_A$$

$$\geqslant \sum_1^n E[s_k^2 I_{A_k} + 2(s_n - s_k)\, s_k I_{A_k}]$$

$$= \sum_1^n [Es_k^2 I_{A_k} + 2E(s_n - s_k) \cdot Es_k I_{A_k}] = \sum_1^n Es_k^2 I_{A_k}$$

$$\geqslant a^2 \sum_1^n EI_{A_k} = a^2 E \sum_1^n I_{A_k} = a^2 EI_A = a^2 P(A).$$

Remark 1 *). If RRV's $\xi_1, \ldots, \xi_n$ are independent, and $E\xi_k = 0$, $E \mid \xi_k \mid^\alpha < \infty \ \forall k$ and certain $\alpha \geqslant 1$, then

$$(1) \qquad P\{\max_k \mid \xi_1 + \ldots + \xi_k \mid \geqslant a\} \leqslant \frac{E \mid \xi_1 \mid^\alpha + \ldots + E \mid \xi_n \mid^\alpha}{a^\alpha}$$

for all $a > 0$. This inequality contains the previous one (when $a = 2$), and is proved in a similar manner. In fact, it follows from $A = \sum_1^n A_k$ that

$$\mid s_n \mid^\alpha \geqslant \mid s_n \mid^\alpha I_A = \sum_1^n \mid s_n \mid^\alpha I_{A_k} = \sum_1^n \mid s_k + (s_n - s_k) \mid^\alpha I_{A_k}.$$

Applying the Jensen inequality to the conditional expectation $E_k = E(./\xi_1, \ldots, \xi_k)$ (see § 21), we obtain

$$E_k \mid s_k + (s_n - s_k) \mid^\alpha I_{A_k} = I_{A_k} E_k \mid s_k + (s_n - s_k) \mid^\alpha$$

$$\geqslant I_{A_k} \mid s_k + E(s_n - s_k) \mid^\alpha = \mid s_k \mid^\alpha I_{A_k} \geqslant a^\alpha I_{A_k}.$$

But then

$$E \mid s_n \mid^\alpha \geqslant \sum_1^n a^\alpha E I_{A_k} = a^\alpha E I_A = a^\alpha P(A).$$

Remark 2. For any RRV's $\xi_1, \ldots, \xi_n$ such that $E \mid \xi_k \mid < \infty \ \forall k$, the inequality is valid

$$P\{\max_{1 \leqslant k \leqslant n} \mid \xi_1 + \ldots + \xi_k \mid \geqslant a\} \leqslant \frac{E \mid \xi_1 \mid + \ldots + \mid \xi_n \mid}{a},$$

where a is any positive number. Certainly, this inequality is rougher than (1), but it has the advantage that it does not require the independence of the RV's $\xi_1, \ldots, \xi_n$ and $E\xi_k = 0 \ \forall k$.

Proof is obtained from the inclusion

$$\{\max_k \mid \xi_1 + \ldots + \xi_k \mid \geqslant a\} \subset \{\mid \xi_1 \mid + \ldots + \mid \xi_n \mid \geqslant a\}$$

and Markov inequality (see § 17 for the case $k = 1$, $\varepsilon = a$, and $\xi = \mid \xi_1 \mid + \ldots + \mid \xi_n \mid$).

§ 20. Moment Inequalities

A. Liapunov inequality. If $0 < \alpha \leqslant \beta$, then

$$\boxed{(E \mid \xi \mid^\alpha)^{1/\alpha} \leqslant (E \mid \xi \mid^\beta)^{1/\beta}}.$$

We certainly assume that the moments involved exist (i.e., $E \mid \xi \mid^\beta < \infty$).

*) The sign * distinguishes harder propositions or problems.

Proof. Apply the Jensen inequality [$\varphi(x) = \mid x \mid^\lambda$]

$$\mid E\eta \mid^\lambda \leqslant E \mid \eta \mid^\lambda$$

to the case $\lambda = \beta/\alpha \geqslant 1$, $\eta = \mid \xi \mid^\alpha$. We obtain

$$(E \mid \xi \mid^\alpha)^{\beta/\alpha} \leqslant E \mid \xi \mid^{\alpha\lambda} = E \mid \xi \mid^\beta.$$

B. Hölder-Minkowski inequality.

$$E \mid \xi\eta \mid \leqslant (E \mid \xi \mid^p)^{1/p} \cdot (E \mid \eta \mid^q)^{1/q}$$

if

$$\frac{1}{p} + \frac{1}{q} = 1, \quad p > 1, \quad q > 1,$$

$$E \mid \xi \mid^p < \infty, \qquad E \mid \eta \mid^q < \infty.$$

Proof. Put

$$\hat{\xi} = \frac{\mid \xi \mid}{(E \mid \xi \mid^p)^{1/p}}, \qquad \hat{\eta} = \frac{\mid \eta \mid}{(E \mid \eta \mid^q)^{1/q}}.$$

It should be shown that

$$E\hat{\xi}\hat{\eta} \leqslant 1.$$

(a) Thanks to the convexity of the function $\varphi(x) = -\ln x$ on $(0, \infty)$, we have

$$\ln(\alpha x + \beta y) \geqslant \alpha \ln x + \beta \ln y$$

for all positive x, y, α, β such that $\alpha + \beta = 1$.

Hence, we get

$$\alpha x + \beta y \geqslant x^\alpha y^\beta$$

for the same numbers.

(b) In particular, if

$$\alpha = \frac{1}{p}, \qquad \beta = \frac{1}{q}, \qquad x = \hat{\xi}^p, \qquad y = \hat{\eta}^q,$$

then

$$\hat{\xi}\hat{\eta} \leqslant \frac{1}{p}\hat{\xi}^p + \frac{1}{q}\hat{\eta}^q.$$

Applying the ME operation E, we obtain

$$E\hat{\xi}\hat{\eta} \leqslant \frac{1}{p}E\hat{\xi}^p + \frac{1}{q}E\hat{\eta}^q = \frac{1}{p} + \frac{1}{q} = 1.$$

Remark. In the case $p = q = 2$, we obtain the Cauchy-Schwarz inequality

$$E \mid \xi\eta \mid \leqslant (E \mid \xi \mid^2)^{1/2} \cdot (E \mid \eta \mid^2)^{1/2}.$$

§ 21. Conditional Expectation

A. In § 4, Ch. 1, the concept of conditional probability relative to an event was defined. We shall now give the definition of the conditional expectation and conditional probability relative to an RV and σ-algebra of events.

Let A be an event, and $P(A) > 0$. If η is an integrable RRV, then, by definition,

$$E(\eta \mid A) = \int_{\Omega} \eta \, dP_A = \frac{1}{P(A)} \int_A \eta \, dP \tag{1}$$

is the **conditional expectation** of the RV η relative to the event A.

We start with a simple case where ξ is an RV assuming values x_1, x_2, . . . with positive probabilities, i.e.,

$$\xi = \sum_{k \geqslant 1} x_k I_{A_k}$$

where $A_k = \{\xi = x_k\}$, $P(A_k) > 0$. Define an RV $E(\eta \mid \xi)$ by the relation

$$E(\eta \mid \xi)(\omega) = \sum_{k \geqslant 1} E(\eta \mid A_k) I_{A_k}(\omega) = \sum_{k \geqslant 1} E(\eta \mid \xi = x_k) I_{\{\xi = x_k\}}(\omega).$$

By definition, the RV $E(\eta \mid \xi)$ is the conditional expectation of the RV η relative to the RV ξ.

If A is an event of the form $A = \sum_{i \in I} A_i$, where $I \subset \{1, 2, \ldots\}$, then, making use of formula (1), we obtain

$$\begin{aligned} \int_A E(\eta \mid \xi) \, dP &= E I_A E(\eta \mid \xi) = E \sum_{k \geqslant 1} E(\eta \mid A_k) I_{AA_k} \\ &= E \sum_{i \in I} E(\eta \mid A_i) I_{A_i} = \sum_{i \in I} E(\eta \mid A_i) E I_{A_i} \\ &= \sum_{i \in I} \int_{A_i} \eta \, dP = \int_A \eta \, dP, \end{aligned}$$

i.e.,

$$\int_A E(\eta \mid \xi) \, dP = \int_A \eta \, dP \tag{2}$$

for any event A of the form $\sum_{i \in I} A_i$. Put $\mathcal{F}_\xi = \xi^{-1}(\mathbf{B}^1)$. Since the σ-algebra $\mathcal{F}_\xi$ coincides with the class of events of the form $\sum_{i \in I} A_i$, equality (2) holds for all $A \in \mathcal{F}_\xi$. This circumstance may be taken as a basis for the definition of $E(\eta \mid \xi)$.

Let (Ω, S, P) be a basic probability space, and ξ an RV with values in a measurable space $(X, \mathcal{A})$. Put $\mathcal{F}_\xi = \xi^{-1}(\mathcal{A})$. It is clear that $\mathcal{F}_\xi \subset S$.

Theorem. For an integrable RRV η, equality (2) determines an RV $E(\eta \mid \xi)$ which is P-a.c. unique, $\mathcal{F}_\xi$-measurable for all $A \in \mathcal{F}_\xi$, and called the **conditional expectation** of the RV η relative to the RV ξ.

In fact, a measure P and σ-finite measure Q are defined on the measurable space $(\Omega, \mathcal{F}_\xi)$ by the equality

$$Q(A) = \int_A \eta \, dP.$$

If $P(A) = 0$, then $Q(A) = 0$. Therefore, by the Radon-Nikodým theorem, there exists a P-a.c. unique, $\mathcal{F}_\xi$-measurable and integrable function φ such that

$$\int_A \varphi \, dP = Q(A) = \int_A \eta \, dP$$

for all $A \in \mathcal{F}_\xi$. It remains to put $E(\eta \mid \xi)(\omega) = \varphi(\omega)$.

In particular, if B is an event, and $\eta = I_B$, then, by definition, the RV

$$P(B \mid \xi) = E(I_B \mid \xi) \tag{3}$$

is the conditional probability of the event B relative to the RV ξ.

It follows from (2) and (3) when $A = \Omega$ that

$$E(E(\eta \mid \xi)) = E\eta,$$
$$EP(B \mid \xi) = P(B).$$

Theorem. If a σ-algebra $\mathcal{F} \subset S$, and η is an integrable RRV, then the equality

$$\int_A E(\eta \mid \mathcal{F}) \, dP = \int_A \eta \, dP$$

defines a P-a.c. unique, $\mathcal{F}$-measurable RV $E(\eta \mid \mathcal{F})$ for all $A \in \mathcal{F}$. The latter RV is called the conditional expectation of the RV η relative to the σ-algebra $\mathcal{F}$.

The proof is performed similarly.

Note that $E(\eta \mid \xi) = E(\eta \mid \mathcal{F}_\xi)$.

B. As it turns out, there exists a measurable function g: $(X, \mathcal{A}) \to (R^1, \mathbf{B}^1)$ such that

$$E(\eta \mid \xi) = g(\xi),$$

which makes it possible to employ the mnemonic notation

$$E(\eta \mid \xi = x) = g(x),$$

and assume, by definition, that $g(x)$ is the ME of the RV η, provided $\xi = x$, and, more precisely, require that $E(\eta \mid \xi)$ is an RV on a probability space $(X, \mathcal{A}, P_\xi)$, where the probability measure P_ξ on $(X, \mathcal{A})$ is defined by the relation $P_\xi = P \cdot \xi^{-1}$, i.e., $P_\xi(A) = P(\xi^{-1}(A))$ for $A \in \mathcal{A}$.

The statement just formulated is proved simply by means of the Radon-Nikodým theorem, integration variable substitution rule, and taking into account the fact that a (σ-finite) measure Q on $(X, \mathcal{A})$ defined by the relation

$$Q(A) = \int\limits_{\xi^{-1}(A)} E(\eta \mid \xi)\, dP = \int\limits_{\xi^{-1}(A)} \eta\, dP, \quad A \in \mathcal{A}$$

is absolutely continuous with respect to the measure P_ξ on $(X, \mathcal{A})$.

Another proof not involving the concept of measure (and thereby the Radon-Nikodým theorem) is based on the following statement (having a wider area of application).

Proposition. Consider two mappings $\xi: \Omega \to X$ and $f: \Omega \to Y$. Denote the set of all subsets of the set X by $\mathcal{P}(X)$. Then the following conditions are equivalent:

C1. $\{f^{-1}(y) : y \in Y\} \subset \xi^{-1}(\mathcal{P}(X))$.

C2. $\forall x \in X\ \exists y \in Y$ such that $\xi^{-1}(x) \subset f^{-1}(y)$.

C3. $\exists g : X \to Y$ such that $f = g(\xi)$, i.e., $f = g \circ \xi$. Besides, if (Ω, S), $(X, \mathcal{A})$, $(Y, \mathcal{B})$ are measurable spaces, $\Omega \xrightarrow{\xi} X \xrightarrow{g} Y$, and $\xi(S) = \mathcal{A}$, then the measurability of the mapping $f = g \circ \xi$ implies that of the mapping g (moreover, if the mapping ξ is measurable, then the mappings f and g are measurable simultaneously).

E.g., if

$$f(\omega) = E(\eta \mid \xi)(\omega), \quad \omega \in \Omega,$$

then, for any $y \in Y = R^1$, $f^{-1}(y) \in \mathcal{F} = \xi^{-1}(\mathcal{A}) \subset \xi^{-1}(\mathcal{P}(X))$ is valid, i.e., C1 holds; therefore, there exists a mapping $g : X \to Y$ such that $f = g(\xi)$. Besides, for the measurable spaces $(\Omega, \mathcal{F})$, $(X, \mathcal{A})$, $(R^1, \mathbf{B}^1)$ and mappings $\xi : \Omega \to X$, $f : \Omega \to R^1$, we have that the mapping $f = g \odot \xi$ is measurable; therefore, by the proposition, the mapping $g : X \to R^1$ is also measurable.

Proof of the proposition. C1 $\Rightarrow$ C2. Let $x \in X$. If, for a certain $y \in Y$, $S_y = \xi^{-1}(x) \cap f^{-1}(y) \neq \varnothing$ is valid, then it follows from $S_y \in \xi^{-1}(\mathcal{P}(X))$ that $S_y = \xi^{-1}(A)$ for a certain $A \subset X$, i.e.,

$$\xi^{-1}(x) \cap f^{-1}(y) = \bigcup_{u \in A} \xi^{-1}(u).$$

And since $\xi^{-1}(x) \cap \xi^{-1}(u) = \varnothing$ if $x \neq u$, we obtain that $\xi^{-1}(x) \cap \cap f^{-1}(y) = \xi^{-1}(x)$, i.e., $\xi^{-1}(x) \subset f^{-1}(y)$.

C2 $\Rightarrow$ C3. Since for $\forall x \exists$ unique (!) $y \in Y$ such that $\xi^{-1}(x) \subset \subset f^{-1}(y)$, the relation

$$g(x) = y \Leftrightarrow \xi^{-1}(x) \subset f^{-1}(y)$$

defines a mapping $g : X \to Y$. We show that $f = g(\xi)$. Let $\omega \in \Omega$, $\xi(\omega) = x$, and $f(\omega) = y$. Since $\omega \in \xi^{-1}(x) \cap f^{-1}(y)$, $\xi^{-1} \cap f^{-1}(y) \neq \varnothing$; hence, $\xi^{-1}(x) \subset f^{-1}(y)$ and $g(x) = y$, i.e., $f(\omega) = g(\xi(\omega))$.

C3 $\Rightarrow$ C1. For $\forall y \in Y$, we have $f^{-1}(y) = \xi^{-1}[g^{-1}(y)] \in \xi^{-1}(\mathcal{P}(X))$.

To prove the second part of the proposition, we make use of the fact that for any mapping $\xi : \Omega \to X$, the relation

$$\xi\xi^{-1}(A) = A \cap \xi(\Omega)$$

is valid for $A \subset X$. In particular, if (as in our case) $\xi(\Omega) = X$, then $\xi\xi^{-1}(A) = A$ for all $A \in \mathcal{P}(X)$, i.e., $\xi\xi^{-1} = id$ is the identity mapping acting in $\mathcal{P}(X)$. Now, it follows from $f = g\xi$ that $f^{-1} = \xi^{-1}g^{-1}$, whence $g^{-1} = \xi f^{-1}$. Finally, $B \in \mathcal{B} \Rightarrow f^{-1}(B) \in S \Rightarrow \xi(f^{-1}(B)) \in \mathcal{A}$ (due to $\xi(S) = \mathcal{A}$) $\Rightarrow g^{-1}(B) \in \mathcal{A}$, i.e., the mapping g is measurable.

C. Denote the space of integrable RRV's on the probability space (Ω, S, P) by $L^1 = L^1(\Omega, S, P)$. For any σ-algebra $\mathcal{F} \subset S$, an operator $\xi \to E(\xi \mid \mathcal{F})$ acting from $L^1(\Omega, S, P)$ into $L^1(\Omega, \mathcal{F}, P)$ is linear and positive. The latter means that $\xi \geqslant 0$ a.c. implies $E(\xi \mid \mathcal{F}) \geqslant 0$ a.c. In particular,

$$|E(\xi \mid \mathcal{F})| \leqslant E(|\xi| \mid \mathcal{F}) \text{ a.c.}$$

We now give the analogues of the Beppo Levi and Lebesgue theorems.

1°. **Beppo Levi theorem.** Let a sequence $\{\xi_n\} \in L^1$ satisfy the condition

$$\xi_n \uparrow \text{a.c.}, \quad E\xi_n \leqslant c = \text{const.}$$

Then there exists $\xi \in L^1$ such that $E(\xi \mid \mathcal{F}) = \lim\limits_{n\to\infty} E(\xi_n \mid \mathcal{F})$ a.c.

2°. **Lebesgue theorem.** Let $\{\xi_n\} \subset L^1$ again, and

$$\xi_n \to \xi \text{ a.c.}, \quad |\xi_n| \leqslant \xi_0 \in L^1 \text{ a.c.}$$

Then $\xi \in L^1$, and $E(\xi \mid \mathcal{F}) = \lim\limits_{n\to\infty} E(\xi_n \mid \mathcal{F})$.

Remark. If $\mathcal{F} = \{\varnothing, \Omega\}$, then $E(\xi \mid \mathcal{F}) = E\xi$. If, however, $\mathcal{F} = S$, then $E(\xi \mid \mathcal{F}) = \xi$ a.c.; therefore, the equality

$$E(\xi \mid \mathcal{F}) = \lim_{n\to\infty} E(\xi_n \mid \mathcal{F}) \text{ a.c.}$$

means, in particular, that $E\xi = \lim\limits_{n\to\infty} E\xi_n$ and $\xi_n \to \xi$ a.c.

Proof of 1°. By the Beppo Levi theorem, there exists $\xi \in L^1$ such that $\xi_n \uparrow \xi$ a.c., and $E\xi_n \to E\xi$. Put

$$\eta_n = E(\xi_n \mid \mathcal{F}), \qquad \eta = E(\xi \mid \mathcal{F}), \qquad \varepsilon_n = \eta - \eta_n.$$

Then $\varepsilon_n \geqslant 0$ a.c., $\varepsilon_n \downarrow$ a.c., and $E\varepsilon_n \geqslant 0$. By the Beppo Levi theorem again, there exists $\varepsilon \in L^1$ (in our case, even $\varepsilon \in L^1(\Omega, \mathcal{F}, P)$) such that $\varepsilon_n \downarrow \varepsilon$, and $E\varepsilon_n \to E\varepsilon$. But $E\varepsilon_n = E(\xi - \xi_n) \to 0$. Therefore, $\varepsilon \geqslant 0$ a.c., and $E\varepsilon = 0$, which implies $\varepsilon = 0$ a.c.; hence, $\eta_n \to \eta$ a.c.

Proof of 2°. Put $\delta_n = \sup_{p \geqslant 0} | \xi_{n+p} - \xi |$. Since $\xi_n \to \xi$ a.c., $\delta_n \downarrow 0$ a.c. We have

$$| E(\xi_n \mid \mathscr{F}) - E(\xi \mid \mathscr{F}) | = | E(\xi_n - \xi \mid \mathscr{F}) | \leqslant E(| \xi_n - \xi | \mid \mathscr{F}) \leqslant E(\delta_n \mid \mathscr{F}).$$

It remains to prove that $E(\delta_n \mid \mathscr{F}) \to 0$ a.c. But this follows from 1°, since $\{\delta_n\} \subset L^1$, $\delta_n \downarrow 0$ a.c., and $E\delta_n \geqslant 0$. Note that the inclusion $\delta_n \in L^1$ follows from $| \delta_n | \leqslant 2\xi_0 \in L^1$ a.c.

§ 22. Conditional Probability

Let $P(A \mid \mathscr{F}) = E(1_A \mid \mathscr{F})$ be the conditional probability of an event $A \in S$ relative to a σ-algebra $\mathscr{F} \subset S$. E.g., $P(A \mid \xi) = P(A \mid \mathscr{F}_\xi)$, where $\mathscr{F}_\xi = \xi^{-1}(\mathscr{A})$. Put

$$P(A \mid \mathscr{F})(\omega) = P_\omega(A).$$

Thereby, a mapping $(\Omega, S) \ni (\omega, A) \mapsto P_\omega(A) \in R^1$ is defined. As it turns out, the mapping $P_\omega : A \to P_\omega(A)$ is not always a probability measure on S for all (or almost all) $\omega \in \Omega$. In this connection, we introduce the following definition.

Definition. A mapping

$$(\Omega, S) \ni (\omega, A) \to Q_\omega(A) \in [0, 1]$$

is called a **regular conditional probability** relative to a σ-algebra $\mathscr{F} \subset S$ if

(1) $\forall \omega \in \Omega$, the mapping $Q_\omega : S \to [0, 1]$ is a probability measure, and

(2) $\forall A \in S$, we have $Q_\omega(A) = P(A \mid \mathscr{F})$ (ω)-a.c.

We now consider two RV's

$$\xi : (\Omega, S) \to (X, \mathscr{A}) \quad \text{and} \quad \eta : (\Omega, S) \to (Y, \mathscr{B}).$$

Definition. A mapping

$$(\Omega, \mathscr{B}) \ni (\omega, \mathscr{B}) \to Q_\omega(B) \in [0, 1]$$

is called a **regular conditional distribution** of an RV η relative to an RV ξ if

(1) $\forall \omega \in \Omega$, the mapping $Q_\omega : \mathscr{B} \to [0, 1]$ is a probability measure, and

(2) $\forall B \in \mathscr{B}$, we have $Q_\omega(B) = P(\eta \in B \mid \mathscr{F}_\xi)(\omega)$ a.c., where, as usual, $P(\eta \in B \mid \mathscr{F}_\xi) = P(\eta \in B \mid \xi) = E(1_{\{\eta \in B\}} \mid \mathscr{F}_\xi)$, $\mathscr{F}_\xi = \xi^{-1}(\mathscr{A})$.

If we take an arbitrary σ-algebra $\mathscr{F} \subset S$ instead of $\mathscr{F}_\xi \subset S$, then a mapping $(\omega, B) \to Q_\omega(B)$ defined in a similar manner is called a regular conditional distribution of the RV η relative to $\mathscr{F}$. In particular, if $\eta(\omega) = \omega \forall \omega \in \Omega$, then we obtain a regular conditional probability relative to $\mathscr{F}$.

First, we indicate sufficient conditions for the existence of a regular conditional distribution, and then list the properties of such a distribution.

Definition. Two measurable spaces $(X, \mathscr{A})$ and $(Y, \mathscr{B})$ are said to be **equivalent** if there exists a bijection $k : X \to Y$ such that $k(\mathscr{A}) = \mathscr{B}$. Note that this concept is symmetric with respect to the pair of spaces, since $k^{-1} : Y \to X$ is a bijection, and $k^{-1}(\mathscr{B}) = \mathscr{A}$. A measurable space $(X, \mathscr{A})$ is called a **Borel space** if it is equivalent to a certain space $(Y, \mathscr{B}) \subset (R^1, \mathbf{B}^1)$, where Y is a Borel set on the straight line R^1, and $\mathscr{B}$ the class of Borel sets contained in Y.

Remark 1. A separable metric space is Borel, which follows from the (complete, separable) space $(R^\infty, \mathbf{B}^\infty)$ equal to the product of a countable number of spaces $(R^1, \mathbf{B}^1)$ being Borel, and any separable metric space homeomorphic to a certain part of the space $(R^\infty, \mathbf{B}^\infty)$.

Theorem 1. Let (Ω, S, P) be a probability space, and $(Y, \mathscr{B})$ a Borel space. Then, for any σ-algebra $\mathscr{F} \subset S$ and RV $\eta : (\Omega, S) \to (Y, B)$, there exists a regular conditional distribution of the RV η relative to $\mathscr{F}$.

Remark 2. In particular, if $\xi : (\Omega, S) \to (X, \mathscr{A})$ is an RV and $\mathscr{F} = \mathscr{F}_\xi = \xi^{-1}(\mathscr{A})$, then there exists a regular conditional distribution of the RV η relative to the RV ξ.

Proof. Assume that the theorem is valid for the case $(Y, \mathscr{B}) = (R^1, \mathbf{B}^1)$. Show that the general case is reduced to the latter. Since $(Y, \mathscr{B})$ is a Borel space, there exists a bijection k of the space $(Y, \mathscr{B})$ onto a certain part of the space $(R^1, \mathbf{B}^1)$. Consider the RV $\hat{\eta} = k(\eta)$. In view of the assumption, there exists a regular conditional distribution $\hat{Q}$ of the RV $\hat{\eta}$ relative to $\mathscr{F}$. Put

$$Q_\omega(B) = \hat{Q}_\omega(k(B)) \text{ for } B \in \mathscr{B}.$$

For each $\omega \in \Omega$, the mapping $Q_\omega : \mathscr{B} \to [0, 1]$ is a probability measure (taking into account that $k(\Sigma B_n) = \Sigma k(B_n)$). Besides, if $B \in \mathscr{B}$, then, a.c.,

$$Q_\omega(B) = \hat{Q}_\omega(k(B)) = P(k(\eta) \in k(B) \mid \mathscr{F})(\omega) = P(\eta \in B \mid \mathscr{F})(\omega).$$

Thereby, Q is a regular conditional distribution of the RV η relative to $\mathscr{F}$.

We shall now assume that $(Y, \mathscr{B}) = (R^1, \mathbf{B}^1)$. The remaining part of the proof will be broken into several items.

A. Let $\mathbf{R}$ be the set of rational numbers. For any $r \in \mathbf{R}$, we define

$$F_\omega(r) = P(\eta < r \mid \mathscr{F})(\omega) = E(1_{\{\eta < r\}} \mid \mathscr{F})(\omega).$$

Show that, for almost all ω, the function F_ω on $\mathbf{R}$ is

(1) nondecreasing,

(2) continuous on the right, and

(3) $F_\omega(-\infty) = 0$, $F_\omega(+\infty) = 1$.

In fact, denote by A_1, A_2, A_3 the sets of those ω for which properties (1), (2) and (3), respectively, are destroyed. It should be shown that $P(A_i) = 0$, $i = 1, 2, 3$.

(A_1) If $r_1 < r_2$ for r_1 and r_2 from $\mathbf{R}$, then $1_{\{\eta<r_1\}} \leqslant 1_{\{\eta<r_2\}}$; therefore, $F_\omega(r_1) \leqslant F_\omega(r_2)$ a.c. For $r_1 < r_2$, we put

$$A(r_1, r_2) = \{\omega : F_\omega(r_1) > F_\omega(r_2)\}.$$

Since

$$P(A(r_1, r_2)) = 0 \text{ and } A_1 = \bigcup_{r_1<r_2} A(r_1, r_2),$$

we have $P(A_1) = 0$.

(A_2) Let $\mathbf{R} \ni r_n \downarrow 0 \in \mathbf{R}$, and $r \in \mathbf{R}$.

Since

$$1_{\{\eta<r-r_n\}} \uparrow 1_{\{\eta<r\}}, \quad F_\omega(r-r_n) \to F_\omega(r) \text{ a.c.}$$

Thus, if $B_r = \{\omega: F_\omega(r - r_n) \to F_\omega(r)\}$, then

$$P(B_r^c) = 0 \text{ and } A_2 = \bigcup_{r\in\mathbf{R}} B_r^c;$$

hence, $P(A_2) = 0$.

(A_3) Since

$$1_{\{\eta<n\}} \uparrow 1_\Omega = 1, \text{ and } 1_{\{\eta<-n\}} \downarrow 1_\varnothing = 0 \text{ a.c. as } n \uparrow \infty,$$

we have

$$F_\omega(+\infty) = 1, \text{ and } F_\omega(-\infty) = 0 \text{ a.c.};$$

therefore, $P(A_3) = 0$.

B. Extend F_ω from $\mathbf{R}$ to R^1 by putting, for $x \in R^1$,

$$F_\omega(x) = \begin{cases} \lim\limits_{\mathbf{R}\ni r\uparrow x} F_\omega(r) & \text{if } \omega \in A = A_1 \cup A_2 \cup A_3 \\ G(x) & \text{if } \omega \in A \end{cases}$$

where G is an arbitrary DF on the straight line R^1. Now, for each ω, the function F_ω on R^1 satisfies conditions (1)-(3) of the previous item. Show that, for any $x \in R^1$,

$$F_\omega(x) = P(\eta < x \mid \mathscr{F})(\omega) \text{ a.c.},$$

which follows from the following;

(1) this statement is valid for $x = r \in R$,

(2) when $\mathbf{R} \ni r \uparrow x$, we have $F_\omega(r) \to F_\omega(x)$ a.c., and

(3) when $\mathbf{R} \ni r \uparrow x$, we have $1_{\{\eta<r\}} \uparrow 1_{\{\eta<x\}}$; therefore, by the Beppo Levi theorem (see § 21),

$$P(\eta < r \mid \mathscr{F})(\omega) \to P(\eta < x \mid \mathscr{F})(\omega) \text{ a.c.}$$

C. For each $\omega \in \Omega$, the DF F_ω determines a unique measure Q_ω on $(R^1, \mathbf{B}^1)$, the equality $Q_\omega(-\infty, x) = F_\omega(x)$ being valid for all $x \in R^1$. To complete the proof of the theorem, it only remains to show that, for each $B \in \mathbf{B}^1$,

(1) $$Q_\omega(B) = P(\eta \in B \mid \mathscr{F})(\omega) \text{ a.c.}$$

Putting $P_\omega(B) = P(\eta \in B \mid \mathscr{F})(\omega)$, we obtain from the Beppo Levi theorem (see § 21) that, for any sequence $\{B_n\}$ of pairwise disjoint sets from $\mathbf{B}^1$,

(2) $$P_\omega(\Sigma B_n) = \Sigma P_\omega(B_n) \text{ a.c.}$$

(unfortunately, the function P_ω on $\mathscr{B}$ is not always countably additive for almost all ω, since the set of those ω for which the previous equality holds depends on the sequence $\{B_n\}$)

Now, let $\mathscr{E}$ be the class of sets $B \in \mathbf{B}^1$, for which the formula $Q_\omega(B) = P_\omega(B)$ is valid a.c. Since for $B = (-\infty, x)$,

$$P_\omega(B) = P(\eta < x \mid \mathscr{F})(\omega) = F_\omega(x) = Q_\omega(B) \text{ a.c.},$$

$\mathscr{E}$ contains the class $\mathscr{E}_0$ of sets of the form $(-\infty, x)$, where $-\infty \leqslant x \leqslant +\infty$. Further, if $A \in \mathscr{E}$, $B \in \mathscr{E}$, and $A \supset B$, then $C = A \setminus B \in \mathscr{E}$, since

$$Q_\omega(C) = Q_\omega(A) - Q_\omega(B) = P_\omega(A) - P_\omega(B) = P_\omega(C) \text{ a.c.}$$

In particular,

$$[a, b) = (-\infty, b) \setminus (-\infty, a) \in \mathscr{E} \quad \text{when } -\infty < a \leqslant b \leqslant \infty.$$

Therefore, $\mathscr{E}$ contains the semi-algebra $\mathscr{E}_1$ of sets of the form

$$(-\infty, b) \text{ and } [a, b), \qquad -\infty < a \leqslant b \leqslant +\infty.$$

But then, according to (2) and additivity of measures Q_ω, the class $\mathscr{E}$ also contains the minimal algebra $\mathscr{E}_2$ containing $\mathscr{E}_1$.

We also notice that the class $\mathscr{E}$ is monotone, i.e., if $\{B_n\} \subset \mathscr{E}$, $B_n \uparrow B$ or $B_n \downarrow B$, then $B \in \mathscr{E}$, which can be obtained from the Beppo Levi theorem (see § 21) by taking into account that then

$$1_{\{\eta\in B_n\}} \uparrow 1_{\{\eta\in B\}} \text{ or } 1_{\{\eta\in B_n\}} \downarrow 1_{\{\eta\in B\}}.$$

Thus, $\mathscr{E}$ is a monotone class, and contains an algebra $\mathscr{E}_2$. Therefore (see Remark 3), the class $\mathscr{E}$ contains the minimal σ-algebra $\mathscr{E}_3$ containing $\mathscr{E}_2$. Since $\mathscr{E}_3 = \mathbf{B}^1$ and $\mathscr{E} \subset \mathbf{B}^1$, $\mathbf{B}^1 = \mathscr{E}_3 \subset \mathscr{E} \subset \mathbf{B}^1$, i.e., $\mathscr{E} = \mathbf{B}^1$, which proves (1).

Remark 3. At the end of the latter item of the proof of the theorem, we have used the following assertion, viz., the minimal σ-algebra Σ containing an algebra $\mathscr{A}$ coincides with the minimal monotone class $\mathscr{M}$ containing the algebra $\mathscr{A}$. Let us prove this statement.

(1) Since a σ-algebra is a monotone class, $\mathscr{M} \subset \Sigma$. It will be proved below that $\mathscr{M}$ is an algebra. And since $\{A_n\} \subset \mathscr{M}$ implies $\mathscr{M} \ni B_n = \bigcup_{m\leqslant n} A_m \uparrow \bigcup A_n \in \mathscr{M}$, the algebra $\mathscr{M}$ is closed under the countable union operation, i.e., $\mathscr{M}$ is a σ-algebra. Therefore, $\Sigma \subset \mathscr{M}$, or $\Sigma = \mathscr{M}$.

(2) To show that the class $\mathscr{M}$ is an algebra, we prove that the class $\mathscr{M}$ is closed under the complementation and finite intersection operations. Put $\mathscr{M}^c = \{A : A \text{ and } A^c \in \mathscr{M}\}$. Since $\mathscr{A} \supset \mathscr{M}^c \subset \mathscr{M}$, and,

due to the relations

$$A_n \uparrow A \Leftrightarrow A_n^c \downarrow A^c, \qquad A_n \downarrow A \Leftrightarrow A_n^c \uparrow A^c,$$

$\mathscr{M}^c$ is a monotone class, we have $\mathscr{M}^c = \mathscr{M}$, i.e., the class $\mathscr{M}$ is closed under the complementation operation. Now, for each $A \in \mathscr{M}$, we define $\mathscr{M}_A = \{B : B \in \mathscr{M}, AB \in \mathscr{M}\}$. The subclass $\mathscr{M}_A$ of the monotone class $\mathscr{M}$ is monotone, since

$$B_n \uparrow B \Rightarrow AB_n \uparrow AB, \qquad B_n \downarrow B \Rightarrow AB_n \downarrow AB.$$

If $A \in \mathscr{A}$, then $\mathscr{A} \subset \mathscr{M}_A \subset \mathscr{M}$, and, due to the monotonicity of the class $\mathscr{M}_A$, we have $\mathscr{M}_A = \mathscr{M}$. Finally, it follows from $A \in \mathscr{M}_B \Leftrightarrow B \in \mathscr{M}_A$ that $A \in \mathscr{A}$ implies $A \in \mathscr{M}_B$ for all $B \in \mathscr{M}$ $(= \mathscr{M}_A)$, i.e., $\mathscr{A} \subset \mathscr{M}_B \subset \mathscr{M}$ for all $B \in \mathscr{M}$. Thus, $\mathscr{M}_B = \mathscr{M}$, for all $B \in \mathscr{M}$, and this means that the class $\mathscr{M}$ is closed under the intersection operation on a finite number of sets.

Theorem 2. Let (Ω, S, P) be a probability space, and $P(A \mid \mathscr{F})$ a regular conditional probability relative to a σ-algebra $\mathscr{F} \subset S$. Then, for each integrable RV ξ,

$$E(\xi \mid \mathscr{F}) = \int_\Omega \xi(\omega)\, P(d\omega \mid \mathscr{F}) \text{ a.c.}$$

Proof. By definition (of a conditional probability in terms of the conditional expectation), this holds in the case where ξ is the indicator of an event. From the linearity of both sides of the equality (as functionals of ξ) follows that the equality is preserved for simple functions. The passage to the limit with respect to monotone increasing sequences of simple functions proves the equality for nonnegative RV's. It now remains to make use of the linearity of both sides of the equality once again.

Remark 4 (*on conditional density*). Recall that an RV $\xi : (\Omega, S) \to (X, \mathscr{A})$ has distribution density $p(x)$ with respect to measure λ if, for each $A \in \mathscr{A}$,

$$P(\xi \in A) = \int_A p(x)\, \lambda(dx).$$

Meanwhile, for each integrable RRV of the form $g(\xi)$, we have

$$Eg(\xi) = \int_X g(x)\, p(x)\, \lambda(dx).$$

Now, let $\zeta = (\xi, \eta)$, where ξ and η are RV's with values in $(X, \mathscr{A})$ and $(Y, \mathscr{B})$, respectively. Assume that the RV $\zeta = (\xi, \eta)$ has density $p(x, y)$ with respect to measure $\lambda \times \mu$, where λ and μ are measures on $(X, \mathscr{A})$ and $(Y, \mathscr{B})$, respectively. Then

(1) the RV's ξ and η have distribution densities (with respect to measures λ and μ, respectively) of the forms

$$p_\xi(x) = \int_Y p(x, y)\, \mu(dy) \text{ and } p_\eta(y) = \int_X p(x, y)\, \lambda(dx);$$

(2) if $p_\xi(x) > 0$ (λ—a.e.), then, for each integrable RRV of the form $g(\eta)$, we have

$$E(g(\eta) \mid \xi) = \int_Y g(y)\, p(y \mid \xi)\, \mu(dy),$$

where

$$p(y \mid x) = \frac{p(x, y)}{p_\xi(x)}.$$

In particular, for $g(\eta) = 1_B(\eta)$, $B \in \mathscr{B}$, we obtain

$$P(\eta \in B \mid \xi) = \int_B p(y \mid \xi)\, \mu(dy) \text{ a.c.},$$

in which case the function $p(y \mid x)$ is called the density of the conditional distribution of the RV η with respect to (hypothesis) $\xi = x$.

Statement (1) follows from the representation

$$P(\xi \in A) = P\{(\xi, \eta) \in A \times Y\} = \int_A \int_Y p(x, y)\, \lambda(dx)\, \mu(dy)$$

$$= \int_A P_\xi(x)\, \lambda(dx)$$

for each $A \in \mathscr{A}$. Similarly, for each $B \in \mathscr{B}$, we have

$$P(\eta \in B) = \int_B p_\eta(y)\, \mu(dy).$$

Let us verify statement (2). According to the definition of a conditional expectation, for any $A \in \mathscr{A}$, we have

$$\int_{\xi^{-1}(A)} E(g(\eta) \mid \xi) dP = \int_{\xi^{-1}(A)} g(\eta)\, dP = Eg(\eta) \cdot 1_A(\xi)$$

$$= \int_{X \times Y} g(y) \cdot 1_A(x) \cdot p(x, y)\, \lambda(dx)\, \mu(dy)$$

$$= \int_A \left[\int_Y g(y) \frac{p(x, y)}{p_\xi(x)}\, \mu(dy) \right] p_\xi(x)\, \lambda(dx)$$

$$= \int_A h(x)\, p_\xi(x)\, \lambda(dx) = \int_{\xi^{-1}(A)} h(\xi)\, dP,$$

where

$$h(x) = \int_Y g(y) \frac{p(x, y)}{p_\xi(x)}\, \mu(dy) = \int_Y g(y)\, p(y \mid x)\, \mu(dy),$$

which gives

$$E(g(\eta) \mid \xi) = h(\xi) \text{ a.c.}$$

Remark 5 (*on the Bayes formula*). With the assumptions of the previous remark, and assuming that

$$p_\xi(x) > 0 \ (\lambda\text{-a.e.}) \quad \text{and} \quad p_\eta(y) > 0 \ (\mu\text{-a.e.}),$$

we have

$$p(x, y) = p_\xi(x)\, p(y|x) = p_\eta(y)\, p(x|y),$$

whence

$$p(y|x) = \frac{p_\eta(y) p(x \mid y)}{p_\xi(x)},$$

or

$$p(y|x) = \frac{p_\eta(y)\, p(x \mid y)}{\int\limits_Y p_\eta(v)\, p(x \mid v)\, \mu(dv)},$$

which is just what is called the **Bayes formula** (generalizing the corresponding formula of § 7, Ch. 1). It involves a distribution determined by density $p_\eta(y)$, and called the ***a priori*** **distribution** of the quantity η, whereas the distribution determined by density $p(y \mid x)$ is called the ***a posteriori*** **distribution** of the quantity η when $\xi = x$.

Supplement: Subjective Probability

Let (Ω, S, P) be a probability space. We define a relation in the class S of events by putting $A \prec B$ if $P(A) \leqslant P(B)$. In the case $A \prec B$ and $B \prec A$, we will assume the events A and B to be **equivalent**, and write $A \sim B$. We will call the relation $\prec$ the **likelihood ratio**, and say in the case $A \prec B$ that the event A is **no more likely** than B.

We call a certain relation $\prec$ in S and probability measure P on the class S of events **consistent** if the relation $\prec$ is a likelihood ratio with respect to P, in which case we will also say that the relation $\prec$ is consistent with probability P, or that probability is consistent with the relation.

Below, we solve the following problem, viz., under what conditions does the relation $\prec$ in S uniquely determine a probability measure P consistent with this ratio?

Remark 1. Such a problem arises due to the fact that it is sometimes difficult in probability models to specify a probability measure P on a class S of events, but it can be indicated which of any two events is more likely. It is asked whether or not a probability measure P can be determined from such data. Probability P uniquely determined from the likelihood ratio between events is said to be **subjective** (since it is assumed that *a priori* information regarding the likelihood ratio between events for a concrete probability model is subjective, i.e., specified by experts).

Remark 2. Recall that the relation $\prec$ in an arbitrary set S (between two elements) is given by a set $M \subset S \times S$, so that $x \prec y$ means that $(x, y) \in M$. Two elements x and y from S are said to be **comparable** if $x \prec y$ and/or $y \prec x$. A relation $\prec$ in S is called a **(partial) ordering** if it is

(1) transitive, i.e., $x \prec y$, $y \prec z$ implies $x \prec z$ and

(2) reflexive, i.e., $x \prec y$, $y \prec z$ is equivalent to $x = y$.

If P is probability on a class S of events, and $\prec$ the likelihood ratio generable by P, then the following is valid.

A1. Any two events A and B from S are comparable, i.e., $A \prec B$ and/or $B \prec A$.

A2. If A_1, A_2, B_1, B_2 are events such that $A_1A_2 = B_1B_2 = \varnothing$, and $A_i \prec B_i$, $i = 1, 2$, then $A_1 \cup A_2 \prec B_1 \cup B_2$. Besides, if $A_1 \nprec B_1$ and/or $A_2 \nprec B_2$, then $A_1 \cup A_2 \nprec B_1 \cup B_2$.

A3. $\varnothing \prec A$ for any event A; besides, $\varnothing \nsim \Omega$.

A4. $A_n \downarrow A = \bigcap\limits_{n \geqslant 1} A_n$, $A_n \succ B$ for $n = 1, 2, \ldots$, where A_n, B are events, implies $A \succ B$.

Remark 3. A relation $\prec$ in S satisfying conditions A1-A4 does not as yet uniquely determine probability P consistent with the ratio. As a matter of fact, if (Ω, S, P) is a discrete probability space or probability space containing atoms (see Problem 79), then a mapping $P : S \to [0, 1]$ does not fill the whole interval $[0, 1]$, i.e., $P(S) \neq [0, 1]$, and the mapping P can be "stirred", preserving the ordering $\prec$ in $P(S)$, so that the obtained mapping of the class S into $[0, 1]$ may be a probability measure again (different from P). More precisely consider a mapping $\gamma : [0, 1] \to [0, 1]$ preserving ordering for elements from $P(S) \subset [0, 1]$, i.e., $x < y$ for $x, y \in P(S)$ implies $\gamma(x) < \gamma(y)$, and such that the mapping $Q = \gamma \circ P : S \to [0, 1]$ is a probability measure. The indicated mapping γ is unique (and equal to x $\forall x$) only in the case where (for a certain probability measure P consistent with the relation $\prec$) $P(S) = [0, 1]$. It certainly needs proof. But it is desirable to formulate this condition, i.e., assumption that $P(S) = [0, 1]$, without involving probability P and only in terms of likelihood ratio.

Definition. An RV ξ has uniform distribution on $[0, 1]$ if, for any two intervals J_1 and $J_2 \subset [0, 1]$, the relation $J_1 \prec J_2$ is equivalent to $|J_1| \leqslant |J_2|$, where $|J|$ is the length of an interval J.

Remark 4. Note that, in this definition, it is not required to specify a probability measure P. What is required is a relation $\prec$ in the class S of events. Meanwhile, an RV ξ is (as usual) a measurable mapping of (Ω, S), into $(R^1, \mathbf{B})$.

Consider another assumption.

A5. There exists an RV uniformly distributed on $[0, 1]$.

Theorem. Let (Ω, S) be a measurable space, and $\prec$ a relation given on the class S, which satisfies A1-A5. Then there exists a unique probability measure P consistent with the relation $\prec$.

Proof.

A. First, verify the following statements.

S1. $A \prec B \prec C \Rightarrow A \prec C$.

S2. $A \subset B \Rightarrow A \prec B$.

S3. $A_n \uparrow A = \bigcup_{n \geqslant 1} A_n$, $A_n \prec B \Rightarrow A \prec B$.

Let us check S1. First, we verify the additivity property of the relation, i.e., if $AC = \varnothing = BC$, then

$$A \prec B \Leftrightarrow A + C \prec B + C$$

(as always, $A + C = A \cup C$ in the case $AC = \varnothing$). In fact, this follows from the implications

$$A \prec B \Rightarrow A + C \prec B + C,$$

and

$$A \succ B \Rightarrow A + C \succ B + C,$$

which are consequences of A2. Note that this property permits us to eliminate and add (to both sides of the relation $\prec$) events inconsistent with the other events in this relation.

Now, let $A \prec B \prec C$. Then, according to the additivity property of the relation, we obtain

$$A \prec B \Rightarrow A \setminus AB \prec B \setminus AB, \text{ i.e., } A \prec B \Rightarrow AB^c \prec A^cB.$$

Similarly, $B \prec C \Rightarrow BC^c \prec B^cC$. Hence, in view of A2, we obtain $AB^c + BC^c \prec A^cB + B^cC$. Therefore,

$$AB^c + (ABC^c + A^cBC^c) \prec (A^cBC + A^cBC^c) + B^cC,$$

or, eliminating A^cBC^c (by the relation additivity property), and adding ABC (by the same property),

$$A = AB^c + ABC^c + ABC \prec A^cBC + ABC + B^cC = C,$$

i.e., $A \prec C$.

We now verify S2. According to $A2$, $A3$, we have

$$A = A + \varnothing \prec A + (B \setminus A) = B.$$

Verification of S3. We shall first see that $A \prec B \Rightarrow A^c \succ B^c$. In fact, if $A \prec B$, and $A^c \prec B^c$, $A^c \not\prec B^c$, then, according to A2,

$$\Omega = A + A^c \prec B + B^c = \Omega \text{ and } \Omega \not\prec \Omega,$$

which is incorrect.

Now, let $A_n \uparrow A = \bigcup A_n$, and $A_n \prec B \, \forall n$. Then $A_n^c \downarrow A^c = \bigcap A_n^c$, and $A_n^c \succ B^c$, which, according to A4, implies $A^c \succ B^c$, i.e., $A \prec B$.

B. Let ξ be an RV uniformly distributed on $[0, 1]$ (see A5). Then (see the definition), putting $E(a, b) = \{a < \xi < b\}$, we have

$$E(a_1, b_1) \prec E(a_2, b_2) \Leftrightarrow b_1 - a_1 \leqslant b_2 - a_2,$$

and

$$E(a, b) \sim E[a, b) \sim E(a, b] \sim E[a, b].$$

S4. For any event A, there exists a unique number $p = P(A)$ such that $A \sim E(0, p)$.

This statement permits us to define a mapping $P: S \to [0, 1]$ which, as we shall see below, is the required probability measure.

Verification of S4. Put

$$M(A) = \{a \in [0, 1] : E(0, a) \prec A\}.$$

$M(A) \neq \varnothing$ since $E(0, 0) = \varnothing \prec A$; therefore, $0 \in M(A)$. Now, put

$$p = P(A) = \sup_a \{a : a \in M(A)\},$$

and show that $E(0, p) \sim A$. In fact, consider a sequence $\{a_n\} \subset M(A)$ such that $a_n \uparrow p$. Then $E(0, a_n) \prec A$, $E(0, a_n) \uparrow E(0, p)$, which implies, due to S3, that

$$E(0, p) \prec A. \tag{1}$$

If $p = 1$, then $E(0, p) \sim E[0, 1] = \Omega \succ A$, which, along with (1) and S1, gives $E(0, p) \sim A$.

If, however, $p < 1$, then, for any $a > p$, it follows from the definition of the number p that $E(0, a) \succ A$; therefore, if $p < a_n \downarrow p$, then $E(0, a_n) \succ A$, $E(0, a_n) \downarrow E(0, p]$. Hence, according to A4, we have $E(0, p] \succ A$, which gives, along with (1), that $E(0, p) \sim E(0, p]$, and S1 yields $E(0, p) \sim A$.

A number p satisfying the condition $E(0, p) \sim A$ is unique, since if, in addition, $E(0, p') \sim A$, then, according to S1, we have $E(0, p) \sim E(0, p')$, which is equivalent to $p = p'$.

B. Now, we verify that the mapping $P : S \to [0, 1]$ defined by the condition

$$A \sim E(0, P(A))$$

is the required probability measure. To this end, it suffices to verify the following statements.

S5. The mapping P is additive, i.e., $P \sum_1^n A_k = \sum_1^n P(A_k)$.

S6. The mapping P is σ-additive.

S7. The mapping P and the ratio $\prec$ are consistent, i.e.,

$$A \prec B \Leftrightarrow P(A) \leqslant P(B).$$

S8. A probability measure consistent with the relation $\prec$ is unique.

Verification of S5. It suffices to consider the case $n = 2$, i.e., verify the formula

$$P(A + B) = P(A) + P(B) \text{ for } AB = \varnothing, \quad A, B \in S.$$

Below, we use the transitivity of the equivalence relation $A_1 \sim A_2 \sim A_3 \Rightarrow A_1 \sim A_3$ (see S1) and uniqueness of $P(A)$ determined from the relation $A \sim E(0, P(A))$. It follows from A2 that

$$AB = \varnothing = CD,\ A \sim C,\ A + B \sim C + D$$

implies $B \sim D$. In particular, for

$$C \sim E(0, P(A)),\ D = E[P(A), P(A + B)),$$

because of

$$A \sim C,\ A + B \sim E(0, P(A + B)) = C + D,$$

this gives

$$B \sim D \sim E(P, (A), P(A + B)) \sim E(0, P(A + B) - P(A));$$

whence, due to $B \sim E(0, P(B))$, we obtain

$$P(A + B) - P(A) = P(B)$$

(note that $P(A) \leqslant P(A + B)$, since $A \subset A + B$).

Verification of S6. It suffices to see that $S \ni A_n \downarrow \varnothing$ implies $P(A_n) \to 0$. In fact, let $P(A_n) \downarrow \varepsilon$. Since $P(A_n) \geqslant \varepsilon$, $A_n \succ E(0, \varepsilon)$; therefore, according to A4,

$$\varnothing = \cap A_n \succ E(0, \varepsilon),$$

which implies $\varepsilon = 0$.

Verification of S7. Since $A \sim E(0, P(A))$, $B \sim E(0, P(B))$, and the relation $\prec$ is transitive (see S1),

$$A \prec B \Leftrightarrow E(0, P(A)) \prec E(0, P(B)) \Leftrightarrow P(A) \leqslant P(B).$$

Verification of S8. Let Q be a probability measure consistent with the relation $\prec$. It suffices to show that

$$Q\{E(0, a)\} = a \quad \forall a \in [0, 1], \tag{2}$$

since then $A \sim E(0, P(A))$ implies

$$Q(A) = Q\{E(0, P(A))\} = P(A),$$

i.e., $Q = P$ on S.

It now remains to verify (2). Put $Q(x) = Q\{E[0, x])\} = Q\{E(0, x)\}$. Since

$$E[0, b) = E[0, a) + E[a, b)$$

(when $0 \leqslant a \leqslant b \leqslant 1$), due to the additivity of measure Q, we have

$$Q[a, b) = Q(b) - Q(a),$$

and because $[a, b) \sim [0, b - a)$, we obtain $Q(b - a) = Q(b) - Q(a)$, i.e., $Q(x + y) = Q(x) + Q(y)$ when $x \geqslant 0$, $y \geqslant 0$, $x + y \leqslant 1$, from which it follows, in view of the positive semi-definiteness (and, hence, monotonicity) of the function Q, and $Q(1) = 1$, that $Q(x) = x$, or formula (2).

Problems

1. Show that $A \supset B$ implies $P(A - B) = P(A) - P(B)$.
2. Prove that

$$P(A \Delta B) = P(A) + P(B) - 2P(AB) = P(A \cup B) - P(AB),$$

where $A \Delta B$ is the symmetric difference of two events A and B, i.e.,

$$A \Delta B = (A \setminus B) \cup (B \setminus A) = (A \cup B) - (AB) = AB^c \cup A^c B.$$

3. Show that

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC).$$

4. Dice are thrown once. Describe the probability space. Describe the event A "the sum of the faces is not less than five". Compute the probability of this event.

5. Dice are thrown twice. Describe the probability space. Describe the following events:

A "the sum of the faces is 10" and B "a six is obtained at least once".

Find the probabilities of the events A, B and AB.

6. Dice are thrown twice. Consider the following events:

A "the sum of the faces is even" and B "the sum of the faces is not less than 10".

Find the probability that only one of the events A or B will occur.

7. Dice are thrown until a six occurs. Describe the probability space. Find the probabilities of the following events:

A "a six is obtained in the first two throws" and B "the number of throws is odd".

8. Three shots were fired at a target. Let A_i be the event "hitting the target on the i-th shot". Express the following events in terms of A_1, A_2 and A_3:

A "there were three hits",

B "there were no hits",

C "there was only one hit"

and

D "there were not less than two hits".

9. Let $\{A_n : n = 1, 2, \ldots\}$ be a sequence of events, and B_m the event "occurrence of only A_m among the events $A_1, \ldots, A_m$. Express B_m in terms of $A_1, \ldots, A_m$. Prove that the events $B_1, B_2, \ldots$ are pairwise inconsistent. Express the event $\bigcup_{n \geqslant 1} B_n$ in terms of $A_1, A_2, \ldots$.

10. A fair coin is tossed until the head occurs twice in a row. Describe the probability space. Find the probability that the number of tosses will be even.

11. A fair coin is tossed until one of the sides occurs twice in a row. Describe the probability space. Find the probability that the number of tosses will be even.

12. Assume that only contacts 1, 2 and 3 may be out of order in an electric circuit (breaking it) (see Fig. 6). The reliability of each contact (i.e., probability that the contact will not be out of order during a time interval under consideration) equals q. Construct the probability space.

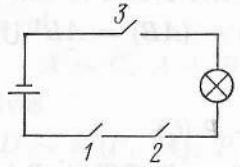

Fig. 6

Consider the events

A_i "the i-th contact is out of order",

A "only one of the contacts is out of order"

and

B "the lamp does not glow".

Find $P_A(A_i)$ and $P_B(A_i)$.

13. Three players a, b, c of the same skill take part in a chess tournament as follows: in the first round, a and b meet, while the player c is free; then, in each subsequent round, the winner of the previous one meets the player who has had a rest; the tournament continues until one of the players wins two games in a row. It is assumed that there are no draws. Construct the probability space. Compute the probability of the events:

A "the winner will be a",

B "the winner will be b",

C "the winner will be c"

and

D "the winner will be unknown until the n-th round".

14. Let P_m be the probability that m events will occur among $A_1, \ldots, A_n$. Show that

$$P_m = \sum_{k=0}^{n-m} (-1)^{k\,m+k} C_m S_{m+k},$$

where

$$S_k = S_k(A_1, \ldots, A_n) = \sum_{1 \leqslant i_1 < \ldots \leqslant i_k < n} P(A_{i_1} \ldots A_{i_k}).$$

15. $n = 3$ people came into a lift. The lift stops at $N = 7$ floors. Find the probability that all the people will go out at different floors.

16. n people are lined in an arbitrary order. Find the probability that two persons A and B will stand side by side. Also, find the probability that there will be r people between A and B.

17. Show that the independence of two events A and B implies that of the events A and B^c, and also of A^c and B^c.

18. Put

$$A^\sigma = \begin{cases} A & \text{if } \sigma = 0 \\ A^c & \text{if } \sigma = 1 \end{cases}.$$

Let n events $A_1, \ldots, A_n$ be independent. Show that the events

$$A_1^{\sigma_1}, \ldots, A_n^{\sigma_n}$$

are also independent (where $\sigma_k = 0$ or $1\ \forall\, k$).

19. Assume that two events A and B_1 are independent, and so are events A and B_2. Show that A and $B_1 \cup B_2$ are independent if and only if A and $B_1 B_2$ are independent.

20. A coin is tossed n times. The event A is "the head is obtained not more than once", and the event B "the head and tail are obtained not less than once each". Show that

when $n = 2$, the events A and B are dependent,

when $n = 3$, the events A and B are independent, and

when $n = 4$, the events A and B are dependent.

21. Give an example of two summable RV's whose product is not summable.

22. Give an example of two RV's ξ and η such that the RV's ξ, η and $\xi \cdot \eta$ are summable, and $E(\xi\eta) \neq E\xi \cdot E\eta$.

23. Assume that n RV's $\xi_1, \ldots, \xi_n$ are independent, and each takes the values $+1$ and -1 with probabilities p and $q = 1 - p$, respectively. Find the distribution of the RV's $\eta_n = \xi_1 \ldots \xi_n$, i.e., evaluate

$$p_n = P(\eta_n = +1) \quad \text{and} \quad q_n = P(\eta_n = -1)$$

(use the equality $E\eta_n = E\xi_1 \ldots E\xi_n$).

24. Let an RV ξ assume a finite number of nonnegative values $x_1, \ldots, x_r$. Prove that

$$\lim_{n\to\infty} \frac{E\xi^{n+1}}{E\xi^n} = \max(x_1, \ldots, x_r),$$

and

$$\lim_{n\to\infty} \sqrt[n]{E\xi^n} = \max(x_1, \ldots, x_r).$$

25. Let ξ be an RV assuming nonnegative integer values, and $E\xi < \infty$. Show that

$$E\xi = \sum_{i \geqslant 1} P(\xi \geqslant i).$$

26. Let an RV ξ assume values $x_1, \ldots, x_m$, and an RV η values $y_1, \ldots, y_n$. Given that the RV ξ^i and η^j are uncorrelated when $i = 1, \ldots, m-1$ and $j = 1, \ldots, n-1$, i.e.,

$$E\xi^i\eta^i = E\xi^i \cdot E\eta^j,$$

show that the RV's ξ and η are independent.

27. Let there n letters be written, with the addresses superscribed at random, and ξ_n the number of letters to be received by the correct addressees. Show that $E\xi_n = 1$.

28. Let ξ and η be two independent RV's assuming values x_1, x_2, ... with probabilities p_1, p_2, ... and q_1, q_2, ..., respectively. Evaluate

$$P(\xi = \eta).$$

29. Let ξ and η be two independent RV's assuming nonnegative integer values, and $E\xi < \infty$. Prove that

$$E \min(\xi, \eta) = \sum_{i \geqslant 1} P(\xi \geqslant i) \cdot P(\eta \geqslant i).$$

30. Two persons toss a coin n times each. Find the probability that everyone will obtain the same number of heads.

31. Show that

(a) $E\xi^2 = 0$ is equivalent to $\xi = 0$ a.c., i.e.,

$$P(\xi = 0) = 1,$$

and

(b) var $\xi = 0$ is equivalent to $\xi = c$ a.c. for a certain number c.

32. Let ξ be an RV, and $E\xi^2 < \infty$. Show that

$$E(\xi - c)^2 \geqslant E(\xi - c_0)^2 = \text{var } \xi$$

for $c_0 = E\xi$ and $\forall c$.

33. Let ξ and η be two RV's with finite second moments. Show that

$$E(\eta - a\xi - b)^2 \geqslant E(\eta - a_0\xi - b_0)^2 = (1 - \rho^2) \text{ var } \eta$$

for all a and b, where

$$a_0 = \frac{\text{cov}(\xi, \eta)}{\text{var } \xi}, \quad b_0 = E\eta - a_0 E\xi,$$

$$\rho = \text{cor}(\xi, \eta),$$

and that if var $\xi = 0$, then $a_0 = 0$.

34. Let ξ and η be two RV's with finite second moments. Show that the following statements are equivalent, viz.,

1°. For certain numbers a and b, $\eta = a\xi + b$ a.c., and

2°. $\rho = \text{cor}(\xi, \eta) = +1$ or -1.

35. (a) Let two RV's ξ_1 and ξ_2 be i.i.d. and have finite second moments. Show that the RV's

$$\eta_1 = \xi_1 + \xi_2 \quad \text{and} \quad \eta_2 = \xi_1 - \xi_2$$

are uncorrelated.

(b) Give an example of uncorrelated, but dependent RV's.

36. Let an RV ξ have the binomial distribution Bi (n, p) with $p = \frac{1}{2}$. Find $E\eta$, where $\eta = \sin \frac{\pi\xi}{2}$.

37. Let an RV ξ assume nonnegative integer values. Show that the following statements are equivalent, viz.,

1°. The RV ξ has a geometric distribution, i.e.,

$$P(\xi = k) = pq^k, \qquad k = 0, 1, 2, \ldots$$

for a certain number $p = 1 - q$, $0 < p < 1$, and

2°. $P(\xi - k = n \mid \xi \geqslant k) = P(\xi = n) \quad \forall k, n.$

38. Given that two RV's ξ_1 and ξ_2 are independent, and have the same geometric distribution, prove that

$$P(\xi_1 = k \mid \xi_1 + \xi_2 = n) = \frac{1}{n+1}, \quad k = 0, 1, \ldots, n.$$

Is the converse statement valid?

39. Given that two RV's ξ_1 and ξ_2 are independent, and have the Poisson distribution with parameters λ_1 and λ_2, respectively, show that

$$P(\xi_1 = k \mid \xi_1 + \xi_2 = n) = {}^nC_k p^k q^{n-k} \quad (k = 0, 1, \ldots, n),$$

where $p = \frac{\lambda_1}{\lambda_1 + \lambda_2} = 1 - q$.

40. (a) Given that an RV ξ has the Poisson distribution with parameter λ, evaluate $E = \frac{1}{1+\xi}$.

(b) Let the generating function of an RV ξ assuming nonnegative integer values be equal to $p(z) = Ez^\xi$, $|z| \leqslant 1$. Show that, when $\alpha > 0$,

$$E \frac{1}{\xi + \alpha} = \int_0^1 z^{\alpha - 1} p(z)\, dz.$$

41. Let ξ_1 and ξ_2 be two independent RV's having the geometric distribution

$$P(\xi_i = k) = p_i q_i^k, \quad i = 1, 2, \quad k = 0, 1, 2, \ldots.$$

Show that the RV

$$\xi = \min(\xi_1, \xi_2)$$

has a geometric distribution. Find a parameter for this distribution.

42. N people speak over N public telephones. The duration of a talk measured in seconds has a geometric distribution with ME μ. Find the average waiting time for one of the booths to be vacated.

43. Let ξ be a random number of factory-made parts. The probability that a part will be defective is p. Denote the number of defective parts by ξ_1, and that of good quality by ξ_2. Show that the RV's ξ_1 and ξ_2 are independent if and only if the RV ξ has the Poisson distribution.

44. Given that two RV's ξ_1 and ξ_2 are independent, assume the values $1, \ldots, N$, and

$$P(\xi_1 = k) = P(\xi_2 = k) = 1/N, \qquad k = 1, \ldots, N.$$

Put

$$U = \min(\xi_1, \xi_2), \quad V = \max(\xi_1, \xi_2), \quad D = V - U.$$

Find the distribution of the RV's U, V and D.

45. Given that $n+1$ RV's $\xi_1, \ldots, \xi_{n+1}$ are independent, and each assumes the values 0 and 1 with probabilities p and $q = 1 - p$, respectively. Put

$$\eta_i = \xi_i + \xi_{i+1} \pmod 2, \quad \gamma = \sum_1^n \eta_i.$$

Find $E\gamma$ and var γ.

46*. Let ξ_1 and ξ_2 be two independent RV's assuming nonnegative values, and $\xi_1 + \xi_2$ have the binomial distribution. Show that each of the variables ξ_1 and ξ_2 has the binomial distribution. Will this statement remain valid if the binomial distribution is replaced by

(a) the negative binomial distribution

or

(b) the Poisson distribution?

47. Given that three independent RV's ξ, η and ζ assume integer values $1, 2, \ldots, n$ with probability $\frac{1}{n}$, find

$$P(\xi + \eta = \zeta),\ P(\xi + \eta = 2\zeta),\ P(\xi + \eta + \zeta = n + 1).$$

48. Let ν be a geometrically distributed RV. Find the generating functions of the RV's

$$\nu_+ = \max(N, \nu),\ \nu_- = \min(N, \nu),$$

where N is a nonnegative integer.

49. Let $\nu, \xi_1, \ldots, \xi_n$ be independent RV's assuming nonnegative integer values, $\xi_1, \ldots, \xi_n$ distributed identically, the generating function of each of them be $p(z)$, $|z| \leqslant 1$, and the RV ν have the binomial distribution Bi (n, p). Put

$$S_0 = 0,\ S_k = \xi_1 + \ldots + \xi_k,\ k = 1, \ldots, n.$$

Find the generating function of the RV S_ν.

50. Let two RV's ξ and η assume nonnegative integer values, and

$$P(\xi = n,\ \eta = k) = \begin{cases} \dfrac{\lambda^n}{n!} e^{-\lambda}\, {}^nC_k p^k (1-p)^{n-k} & \text{when } 0 \leqslant k \leqslant n \\ 0 \text{ when } k > n \end{cases},$$

where $\lambda > 0$, $0 \leqslant p \leqslant 1$. Find the joint generating function of the RV's ξ and η, i.e.,

$$P(z_1, z_2) = E z_1^\xi \cdot z_2^\eta,\ |z_1| \leqslant 1,\quad |z_2| \leqslant 1,$$

and cor (ξ, η).

51. Show that the generating function of the number of inversions ξ in a random permutation of n numbers (all n! permutations being equally likely) is

$$P(z) = \sum_{i=0}^{{}^nC_2} P(\xi = i)\, z^i = \prod_{k=1}^n \frac{1 - z^k}{1 - z} \cdot \frac{1}{n!}$$

Find $E\xi$ and var ξ.

52*. Let an RV ξ have the negative binomial distribution $\overline{\text{Bi}}$ (n, p), and

$$\alpha_k = P(\xi = k), \quad \beta_k = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots,$$

$$\lambda = n\frac{q}{p}, \quad \delta = n\frac{q^2}{p^2}, \quad q \leqslant 1/4.$$

Then

$$(*) \qquad |\alpha_k - \beta_k| \leqslant 2\delta \ \forall\ k.$$

Moreover, if $\xi_1, \ldots, \xi_n$ are independent RV's, the RV ξ_s has the geometric distribution

$$P(\xi_s = k) = p_s q_s^k, \quad q_s = 1 - p_s,$$

with

$$\xi = \xi_1 + \ldots + \xi_n,$$

$$\lambda = E\xi = \sum_1^n E\xi_s = \sum_1^n \frac{q_s}{p_s}, \quad \delta = \sum_1^n \frac{q_s^2}{p_s^2}, \quad q_s \leqslant 1/4\ \forall\ s,$$

then $(*)$ is fulfilled.

53. Let A and B be two events, and 1_A and 1_B their indicators. Show that

(1) var $1_A = P(A) \cdot P(A^c)$,

(2) cov $(1_A, 1_B) = P(AB) - P(A) \cdot P(B)$,

and

(3) if $0 < P(A) < 1$, $0 < P(B) < 1$, then the independence of the events A and B is equivalent to cor $(1_A, 1_B) = 0$.

54. Applying the ME operation to both sides of the equalities

$$1 - 1_{\bigcup_1^n A_k} = 1_{(\bigcup_1^n A_k)^c} = 1_{\bigcap_1^n A_k^c} = \prod_1^n 1_{A_k^c} = (1 - 1_{A_1}) \ldots (1 - 1_{A_n}),$$

obtain the formula for the probability of the union of events in terms of the probabilities of their intersections.

55. Let ξ be an RV and $0 < \text{var}\ \xi < \infty$. Show that

$$P\left\{-3.2 < \frac{\xi - E\xi}{\sqrt{\text{var}\ \xi}} < 3.2\right\} > 0.9.$$

56. Let ξ be an RV, and

$$a = E\xi, \quad 0 < b = [E(\xi - a)^{10}]^{\frac{1}{10}} < +\infty.$$

Show that

$$P\left\{-2 < \frac{\xi - a}{b} < 2\right\} > 0.999.$$

57. Consider a sequence of i.i.d. RV's $\xi_1, \xi_2, \ldots$. Denote the random number of jumps in neighbouring elements of the sequence

$\xi_1, \ldots, \xi_{n+1}$ by $J(n) = \sum_1^n 1_{\{\xi_k \neq \xi_{k+1}\}}$. Show that

$$E\left|\frac{J(n)}{n} - p\right|^2 \to 0 \text{ as } n \to \infty,$$

where $p = P(\xi_1 \neq \xi_2)$.

58. Let $\xi_1, \ldots, \xi_n$ be n RV's (possibly, dependent), and $\varepsilon > 0$. Show that

$$P\left\{\left|\sum_1^n \xi_i\right| \geqslant \varepsilon\right\} \leqslant \frac{\sum_1^n E|\xi_i|}{\varepsilon}.$$

59. Let ξ be an RV, f and g numerical functions on R^1, and $A \subset R^1$. Assume that

(1) $f(x) \leqslant 0$ if $x \in A$, $f(x) \leqslant 1$ if $x \in A$,
(2) $g(x) \geqslant 0$ if $x \in A$, $g(x) \geqslant 1$ if $x \in A$,
(3) $f(\xi)$ and $g(\xi)$ are summable RV's, and
(4) $\{\xi \in A\}$ is an event.

Show that

$$Ef(\xi) \leqslant P(\xi \in A) \leqslant Eg(\xi).$$

Hint: $f(\xi) \leqslant 1_{\{\xi \in A\}} \leqslant g(\xi)$.

60. Consider $n+1$ independent RV's $\xi_1, \ldots, \xi_{n+1}$ for each n, assuming the values 1 and 0 with probabilities p and $q = 1 - p$, respectively. Put

$$\eta_i = \xi_1 + \xi_{i+1} \pmod 2, \quad S_n = \sum_1^n \eta_i.$$

Show that, for any $\varepsilon > 0$,

$$P_n = P\left\{\left|\frac{S_n}{n} - E\eta_1\right| \geqslant \varepsilon\right\} \to 0 \text{ as } n \to \infty.$$

61. Let $\Omega = (0, \infty)$, S the class of all subsets of the set Ω, and P a numerical function on S, defined by the formula

$$P(A) = \sum_{k \in A \cap N} 2^{-k} \text{ for } A \in S,$$

where $N = \{1, 2, \ldots\}$. Is (Ω, S, P) a probability space?

62. Let $(X, \mathcal{A}, P)$ be a probability space, and $(Y, \mathcal{B}, Q)$ a triplet constructed by means of a bijective mapping $f: X \to Y$, where $\mathcal{B} = f(\mathcal{A}) = \{f(A) : A \in \mathcal{A}\}$, and Q a numerical function on $\mathcal{B}$, defined by the relation $Q(B) = P(f^{-1}(B))$ for $B \in \mathcal{B}$. Verify that the triplet $(Y, \mathcal{B}, Q)$ is a probability space.

63. Let $\Omega = [a, b]$. Show that the class

$$S = \{[\alpha, \beta) : a \leqslant \alpha \leqslant \beta \leqslant b\}$$

of subsets of the set Ω is a semi-algebra, but not an algebra.

64. Let a class K of subsets of a set Ω consist of one element $A \subset \Omega$. Find the minimal algebra containing K.

65. Let $K = \{A_1, \ldots, A_n\}$ be a decomposition of Ω, i.e.,

(1) the sets $A_1, \ldots, A_n$ pairwise disjoint, and

(2) $\bigcup_1^n A_i = \Omega$.

Prove that the minimal algebra containing K is the class of all unions of the sets $\varnothing = A_0, A_1, \ldots, A_n$.

66*. Denote by S the class of sets of the form $F \cap G$, where F is a closed, and G an open set from R^1. Show that S is a semi-algebra, but not an algebra.

67. Let K be an arbitrary class of subsets of a set Ω. Forming consecutively

(1) a class K_1 consisting of $\varnothing$, Ω and $A \subset \Omega$ such that either $A \in K$ or $A^c \in K$,
(2) a class K_2 of finite intersections of sets from K_1, and
(3) a class K_3 of finite unions of sets from K_2, show that K_3 is the minimal algebra containing K.

68. Let $\{A_n : n = 1, 2, \ldots\}$ be a sequence of subsets of Ω. If $\overline{\lim} A_n$ is the set of elements $\omega \in \Omega$ belonging to the infinite number of sets A_n, and $\underline{\lim} A_n$ the set of elements $\omega \in \Omega$ belonging to all the sets A_n, possibly, except a finite number of these sets, show that

(1) $\underline{\lim} A_n \subset \overline{\lim} A_n$,

(2) $\underline{\lim} A_n = \bigcup_{n \geqslant 1} \bigcap_{k \geqslant n} A_k$, $\overline{\lim} A_n = \bigcap_{n \geqslant 1} \bigcup_{k \geqslant n} A_k$,

(3) $(\underline{\lim} A_n)^c = \overline{\lim} A_n^c$,

(4) $\overline{\lim}(A_n \cup B_n) = \overline{\lim} A_n \cup \overline{\lim} B_n)$, and

(5) if

$$A_n = \begin{cases} A \text{ for an even } n, \\ B \text{ for an odd } n, \end{cases}$$

then

$$\overline{\lim} A_n = A \cup B, \quad \underline{\lim} A_n = A \cap B.$$

69. Let $\{A_n : n = 1, 2, \ldots\}$ be a sequence of subsets of Ω. We will say that there exists $\lim A_n$ if $\underline{\lim} A_n = \overline{\lim} A_n$, in which case

$$\lim A_n = \underline{\lim} A_n = \overline{\lim} A_n.$$

Show that

(1) if $A_1 \supset A_2 \supset \ldots$, then $\lim A_n = \bigcap A_n$,
(2) if $A_1 \subset A_2 \subset \ldots$, then $\lim A_n = \bigcup A_n$, and
(3) if the sets A_n are pairwise disjoint, then $\lim A_n = \varnothing$.

70. Let (Ω, S, P) be a probability space, $\{A_n\} \subset S$, and there exist $\lim A_n$. Show that

$$P(\lim A_n) = \lim P(A_n).$$

71. Let (X, d) be a separable metric space (with a metric d), and $\mathbf{B}$ the collection of Borel sets in this space (i.e., the minimal σ-algebra containing all open balls in this space). A mapping $\xi : \Omega \to X$ is called an RV (with values in the metric space) if

$$\{\xi \in B\} = \{\omega : \xi(\omega) \in B\} \in S$$

for all $B \in \mathbf{B}$.

Show that a mapping $\xi : \Omega \to X$ is an RV if and only if

$$\{\xi \in E\} \in S$$

for all open balls E.

72. Let ξ be an RV with values in a metric space (X, d), and φ a Borel mapping of the metric space (X, d) into a metric space (Y, ρ) (i.e., a mapping $\varphi: X \to Y$ for which the inverse image of a Borel set of elements from Y is a Borel set of elements from X). Then $\eta = \varphi(\xi)$ is an RV with values in (Y, ρ).

73. Let ξ_1 and ξ_2 be two RV's with values in a metric space (X, d). Show that $d(\xi_1, \xi_2)$ is an RV.

74. Let ξ and η be two RV's. Show that the set

$$\{\xi = \eta\} = \{\omega: \xi(\omega) = (\omega)\}$$

is an event.

75. Let $\xi, \xi_1, \xi_2, \ldots$ be a sequence of RV's. Show that each of the sets

$$\{\lim \xi_n = \xi\} = \{\omega: \lim \xi_n(\omega) = \xi(\omega)\}$$

and

$$\{\underline{\lim}\, \xi_n = \overline{\lim}\, \xi_n\} = \{\omega: \underline{\lim}\, \xi_n(\omega) = \overline{\lim}\, \xi_n(\omega)\}$$

is an event.

76. Let $\nu, \xi_1, \ldots, \xi_n$ be $n+1$ RV's on a common probability space, and the RV ν assume the values $0, 1, \ldots, n$ with probability 1. Put

$$S_0 = 0, \quad S_k = \xi_1 + \ldots + \xi_k, \quad 1 \leqslant k \leqslant n.$$

Show that S_ν is an RV.

77. Let $\eta(\omega) = \xi(\omega)^2$ be an RV. Show that $\xi(\omega)$ is not necessarily an RV (give an example).

78*. Let (Ω, S, P) be a probability space. Prove that the set $\{P(A): A \in S\}$ is closed.

79. A set $A \in S$ is called an atom of a probability space (Ω, S, P) if (a) $P(A) > 0$, (b) $S \ni B \subset A$ implies either $P(B) = 0$ or $P(A \setminus B) = 0$. Prove the following statements.

1°. If A and B are atoms, then either $P(A \Delta B) = 0$ or $P(AB) = 0$.

2°. The set of different atoms is at most countable (two atoms A and B coincide by definition if $P(A \Delta B) = 0$).

3°. There exists a decomposition $\Omega = A_0 + \sum_{i \geqslant 1} A_i$ into disjoint sets $\{A_0, A_1, \ldots\} \subset S$, in which $A_1, A_2, \ldots$ are atoms, and the set A_0 does not contain atoms. (A_0 is called the atomless part of the probability space, whereas $\sum_{i \geqslant 1} A_i = \Omega \setminus A_0$ atomic. If $P(A_0) = 0$, then the probability space is said to be atomic.)

4°*. For any number p such that $0 \leqslant p \leqslant P(A_0)$, there exists $A \in S$, and $P(A) = p$.

80. If A is an atom, and ξ an RV, then there exists a number c such that

$$P(\xi = c \mid A) = 1$$

(i.e., the RV assumes constant values on each atom up to a set of zero probability).

81A. Suppose that n RV's $\xi_1, \ldots, \xi_n$ are independent, and each has the uniform distribution on $(0, 1)$. Put

$$U = \min_i \xi_i, \qquad V = \max_i \xi_i$$

and

$$D = V - U, \quad R = U/V.$$

If ξ is an RV, then we will denote its CF and density by F_ξ and f_ξ, respectively. Show that

(a) $F_V(x) = x^n$, $f_V(x) = nx^{n-1}$, $0 < x < 1$,

$$EV = \frac{n}{n+1}, \quad \operatorname{var} V = \frac{n}{(n+2)(n+1)^2},$$

(b) $F_U(x) = 1 - (1-x)^n$, $\quad f_U(x) = n(1-x)^{n-1}$, $\quad 0 < x < 1$,

$$EU = \frac{1}{n+1}, \quad \operatorname{var} U = \operatorname{var} V,$$

(c) $F_{UV}(u, v) = v^n - (v-u)^n$,

$f_{UV}(u, v) = n(n-1)(v-u)^{n-2}$, $\quad 0 < u \leqslant v < 1$,

(d) $F_D(x) = nx^{n-1} - (n-1)x^n$, $\quad f_D(x) = n(n-1)(x^{n-2} - x^{n-1})$,

$n \geqslant 2$, $\quad 0 < x < 1$,

$$ED = \frac{n-1}{n+1}, \quad \operatorname{var} D = \frac{2(n-1)}{(n+1)^2(n+2)},$$

(e) $\operatorname{cor}(U, V) = 1/n$,

and

(f) the RV's R and $\min(\xi_1, \ldots, \xi_{n-1})$ are distributed identically.

81. Suppose that RV's $\xi_1, \xi_2, \ldots$ are independent, and each is distributed in $(0, 1)$ uniformly. Put

$$U_n = \min(\xi_1, \ldots, \xi_n), \qquad V_n = \max(\xi_1, \ldots, \xi_n),$$

and

$$D_n = V_n - U_n, \qquad R_n = U_n/V_n.$$

Show that, as $n \to \infty$, for any $x > 0$,

$$\lim P\{nU_n < x\} = 1 - e^{-x},$$

$$\lim P\{n(1 - V_n) < x\} = 1 - e^{-x},$$

$$\lim P\{n[1 - D_n] < x\} = 1 - (1 + x)e^{-x},$$

and

$$\lim P\{nR_n < x\} = 1 - e^{-x}.$$

82. Two persons A and B are to meet between 20 and 21 hours, having agreed to wait for not more than ten minutes. Suppose that the moments of their arrival at the appointed place are independent, and distributed uniformly between 20 and 21 hours. Find the probability that they will meet.

83. Let ξ and η be two RV's. Denote the densities of the DF's of the RV's ξ and η by f and g, respectively. Given f, find g if (1) $\eta = \xi + c$, (2) $\eta = \sigma\xi + \mu$, $\sigma > 0$, (3) $\eta = \sigma\xi + \mu$, $\sigma < 0$, (4) $\eta = \sigma\xi + \mu$, $\sigma \neq 0$, (5) $\eta = \xi^3$, (6) $\eta = |\xi|$, and (7) $\eta = \xi^2$.

84. Suppose that two RV's ξ_1 and ξ_2 are independent, and each has the exponential distribution with parameter a, i.e.,

$$P(\xi_i < x) = 1 - e^{-ax} \quad \forall x \geqslant 0.$$

Show that

(a) the RV $\xi_1/(\xi_1 + \xi_2)$ is distributed uniformly on (0, 1), and

(b) the RV's

$$\frac{\xi_1}{\xi_1 + \xi_2} \quad \text{and} \quad \xi_1 + \xi_2$$

are independent.

85. Suppose that two RV's ξ_1 and ξ_2 are independent, and each has the normal distribution $N(0, 1)$. Show that the RV

$$\eta = \frac{1}{2}(\xi_1^2 + \xi_2^2)$$

has the exponential distribution with parameter $a = 1$, i.e.,

$$P(\eta < x) = 1 - e^{-x} \quad \forall x \geqslant 0.$$

86. Let the DF F of an RV ξ be continuous. Show that the RV $\eta = F(\xi)$ is distributed uniformly on (0, 1).

87. Let two RV's ξ and η be independent, and distributed uniformly on [0, 1]. Find the probability that the roots of the quadratic equations

(a) $x^2 + \xi x + \eta = 0$,

(b) $x^2 + Vx + U = 0$, where

$$U = \min(\xi, \eta), \quad V = \max(\xi, \eta),$$

will be real.

88. Suppose that two RV's ξ_1 and ξ_2 are independent, and each has the normal distribution $N(\mu, \sigma^2)$. Show that

$$E \max(\xi_1, \xi_2) = \mu + \frac{\sigma}{\sqrt{\pi}}, \quad E \min(\xi_1, \xi_2) = \mu - \frac{\sigma}{\sqrt{\pi}}.$$

89. A point $x \in R^1$ is a point of increase for a DF F if, for any $\varepsilon > 0$,

$$F(x + \varepsilon) - F(x - \varepsilon) > 0.$$

Give an example of the DF of a discrete RV for which each point $x \in R^1$ is a point of increase.

90*. Elements a_i and a_{i+1} in a permutation $(a_1, \ldots, a_n)$ "form ordering" if $a_i < a_{i+1}$. It is assumed that the element a_1 forms ordering with all the other elements, and that all the numbers $a_1, \ldots, a_n$ are different. Denote the number of points of increase in a random permutation (all the $n!$ permutations are equally likely) by η_n. We also assume that the RV's $\xi_1, \ldots, \xi_n$ are independent, and distributed on (0, 1) uniformly. Show that

(a) $$P(\xi_1 + \ldots + \xi_n < x) = \frac{1}{n!}\sum_{j=0}^{[x]} (-1)^j \, {}^nC_j (x - j)^n \quad \forall x > 0, \text{ and}$$

(b) $$P(\eta_n \leqslant k) = P(\xi_1 + \ldots + \xi_n < k) \quad \forall k = 1, 2, \ldots, n.$$

91. Let two RV's ξ_1 and ξ_2 have the joint normal distribution. Show that the RV's ξ_1 and ξ_2 are independent if and only if

$$\operatorname{cov}(\xi_1, \xi_2) = 0.$$

92. Let two RV's ξ and η be i.i.d. normally. Show that the RV's $\xi + \eta$ and $\xi - \eta$ are independent.

93. Let r RV's $\xi_1, \ldots, \xi_r$ have the joint normal distribution $N_r(\mu, A)$. Show that the RV's

$$\xi = \sum_1^r a_i \xi_i \quad \text{and} \quad \eta = \sum_1^r b_j \xi_j$$

are independent if and only if $(Aa, b) = 0$, where

$$a = (a_1, \ldots, a_r), \quad b = (b_1, \ldots, b_r).$$

In particular, if, moreover, the RV's $\xi_1, \ldots, \xi_r$ are i.i.d. normally, then the independence of the RV's ξ and η is equivalent to

$$(a, b) = \sum_1^r a_i b_i = 0.$$

94. Let a random vector $\xi = (\xi_1, \ldots, \xi_r)$ have the normal distribution $N_r(\mu, A)$. Put

$$\rho_{ij} = \operatorname{cor}(\xi_i, \xi_j), \quad R = \{\rho_{ij}\}, \quad A = \{a_{ij}\}.$$

Show that

(a) $$\rho_{ij} = \frac{a_{ij}}{\sqrt{a_{ii}a_{ij}}},$$

and

(b) $$R = \sum^{-1} A \sum^{-1},$$

where

$$\sum = \begin{vmatrix} \sigma_1 & 0 & \vdots & 0 \\ 0 & \sigma_2 & \vdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \vdots & \sigma_r \end{vmatrix}, \quad \sigma_i > 0, \ \sigma_i^2 = \operatorname{var} \xi_i = a_{ii}.$$

95. Let f_1 and f_2 be two densities of two-dimensional normal distributions with zero ME's and unit variances, but different correlation coefficients. Prove that (a) the function $\frac{1}{2}(f_1 + f_2)$ is the distribution density of a certain random vector (ξ_1, ξ_2),

(b) the distribution of the vector (ξ_1, ξ_2) is not normal, and

(c) each of the variables ξ_1 and ξ_2 has the normal distribution $N(0, 1)$.

96. Let a random vector (ξ_1, ξ_2) have the normal distribution $E\xi_1 = E\xi_2 = 0$, $\operatorname{var} \xi_1 = \operatorname{var} \xi_2 = 1$, $E\xi_1\xi_2 = \rho$. Show that

$$P_i'(\xi_1\xi_2 > 0) = \frac{1}{2} + \frac{1}{\pi} \sin^{-1}\rho \qquad P(\xi_1\xi_2 < 0) = \frac{1}{\pi} \cos^{-1}\rho.$$

97. (*continued*). Prove that $E \max(\xi_1, \xi_2) = \left(\frac{1-\rho}{\pi}\right)^{1/2}$.

98. Let a random vector (ξ, η) have the normal distribution, and $E\xi = E\eta = 0$, $E\xi\eta < 0$. Show that

$$P(\xi \leqslant a, \ \eta \geqslant b) \leqslant P(\xi \geqslant a) \cdot P(\eta \geqslant b)$$

when $a > 0$, $b > 0$.

99. Let two RV's ξ_1 and ξ_2 be independent, and distributed uniformly in $(0, 1)$. Show that the RV's $\eta_1 = \rho \cos \varphi$ and $\eta_2 = \rho \sin \varphi$, where $\rho = \sqrt{-2 \ln \xi_1}$, $\varphi = 2\pi\xi_2$, are independent, and each has the normal distribution $N(0, 1)$.

100. Let two RV's ξ_1 and ξ_2 be independent, and have the normal distributions $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$, respectively. Show that the RV

$$\xi = \frac{\xi_1\xi_2}{\sqrt{\xi_1^2 + \xi_2^2}}$$

has the normal distribution $N(0, \sigma^2)$, where

$$\frac{1}{\sigma} = \frac{1}{\sigma_1} + \frac{1}{\sigma_2}, \quad \sigma_1 > 0, \sigma_2 > 0.$$

Part 2

Stochastic Processes

Introduction: Definition of Random Process. Kolmogorov Theorem on Construction of Processes from Partial Distributions

1°. Examples of random processes are given, e.g., by atmospheric precipitation, the water-level in a river, air temperature relative to time and place, the number of ships to be unloaded, the waiting time interval for the last person in a queue to be served, the number of engaged communication lines relative to time and the velocity of water flow in the ocean relative to depth, place and time.

D1. A set $\xi = \{\xi(t): t \in T\}$ of RV's $\xi(t)$ defined on the same probability space (Ω, S, P) is called a **random function** (RF). If the parameter t plays the role of time, i.e., $T \subset R^1$, then such a random function is more often called a **random process**. If, however, $T \subset \{0, \pm 1, \pm 2, \ldots\}$ or T is an ordered countable set, then a **random sequence** is also spoken of.

Thus, a random function ξ is a mapping $(t, \omega) \to \xi(t, \omega)$ measurable in the second argument.

E.g., $\xi(t)$ may be a Fourier series or polynomial of t with random coefficients. A sequence $\{\xi_k : k = 1, 2, \ldots\}$ of RV's is also a random process (if we put $T = \{1, 2, \ldots\}$, and $\xi(t) = \xi_t$ for $t \in T$).

2°. For a fixed $\omega \in \Omega$, a function $\xi(t, \omega)$ of t is called a (sample) **trajectory** (or function) of the process ξ. As it turns out, a σ-algebra of subsets and probability measure can be introduced on the set of all trajectories, so that the trajectories become elementary events in the probability space constructed. Accordingly, we give two examples.

Example 1. Let (Ω^*, S^*, P^*) be a probability space, $(X, \mathcal{F})$ a measurable space, and T a certain set. Suppose that Ω^* is a function space *alias* a certain class of functions $\omega^*: T \to X$, and the σ-algebra S^* contains sets of the form

$$\{\omega^* : \omega^* \in \Omega^*, \ \omega^*(t) \in B\}, \ t \in T, \ B \in \mathcal{F}.$$

Then the mapping $\xi : T \times \Omega^* \to X$ defined by the formula $\xi(t, \omega^*) = \omega^*(t)$ is a random process. In fact, it only should be verified

that, for any $t \in T$, the function $\xi(t) : \omega^* \to \xi(t, \omega^*)$ is measurable (is an RV), which follows from

$$\{\xi(t) \in B\} = \{\omega^* : \xi(t, \omega^*) \in B\} = \{\omega^* : \omega^*(t) \in B\} \in S$$

for any $B \in \mathcal{F}$. Note that in this example an elementary event ω^* is a trajectory of the process ξ.

Example 2. Let $\xi = \{\xi(t) : t \in T\}$ be a random process on a probability space (Ω, S, P) with values in a measurable space $(X, \mathcal{F})$. Denote the set of all trajectories of the process ξ by Ω^*, and assume that to different $\omega \in \Omega$, there correspond different trajectories. Define a mapping $k : \Omega \to \Omega^*$ associating an elementary event $\omega \in \Omega$ with a trajectory $k(\omega) = \omega^* \in \Omega^*$ (thereby, $k(\omega)(t) = \xi(t, \omega)$). The mapping k is a bijection, in which case the σ-algebra S and probability P on S are transformed into a σ-algebra S^* and probability P^* on S^*, respectively, viz.,

$$S^* = k(S) = \{k(A) : A \in S\},$$

$P^*(A^*) = P(A)$ for $A^* = k(A)$ (i.e., $P^* = k(P) = P \cdot k^{-1}$). The triplet (Ω^*, S^*, P^*) is a probability space in which the trajectories of the initial random process ξ serve as elementary events.

D2. Two processes $\xi = \{\xi(t) : t \in T\}$ and $\eta = \{\eta(t) : t \in T\}$ with values in $(X, \mathcal{F})$, and defined on the same probability space (Ω, S, P), are said to be (stochastically) **equivalent in the wide sense** if

$$P\{\xi(t_1) \in B_1, \ldots, \xi(t_n) \in B_n\} = P\{\eta(t_1) \in B_1, \ldots, \eta(t_n) \in B_n\}$$

for any integer $n \geqslant 1$, $t_k \in T$ and $B_k \in \mathcal{F}$ when $k = 1, \ldots, n$, in which case the processes ξ and η are said to have the same finite-dimensional distributions.

It is assumed that processes having the same finite-dimensional distributions are practically indistinguishable.

The theorem of A. N. Kolmogorov, given below, means that a random process with the same finite-dimensional distributions can be constructed from a family of finite-dimensional distributions made consistent in a natural manner.

3°. Let T be a certain set, and (X_t, d_t) a metric space with metric d_t for $t \in T$. The direct product of a finite number of metric spaces (X_t, d_t), $t \in M$, is a metric space (X_M, d_M), where

$$X_M = \prod_{t \in M} X_t, \quad d_M = \max_{t \in M} d_t.$$

Denote the σ-algebra of Borel sets in the space X_M by $\mathbf{B}_M$, and assume that, for each finite set $M \subset T$, a probability measure P_M on $(X_M, \mathbf{B}_M)$ is given. Consider the **consistency condition** (for the family of measures P_M, $M \subset T$ and $|M| < \infty$), viz.,

$$P_M(A \times X_{M \setminus N}) = P_N(A) \text{ for } N \subset M \text{ and } A \in B_N,$$

in which case the measure P_N is called the **projection** of the measure P_M onto $(X_N, \mathbf{B}_N)$.

We also define a measurable space $(X_T, \mathbf{B}_T)$ in which $X_T = \prod_{t \in T} X_t$, the set $A \times X_{T \setminus M} \subset X_T$ being called a **cylinder** (with base A) for a finite set $M \subset T$ and $A \in \mathbf{B}_M$, and $\mathbf{B}_T$ being the smallest σ-algebra containing all the cylinders.

Theorem. Let, for each $t \in T$, X_t be a Polish space (complete, separable and metric space), and a family of probability measures P_M on $(X_M, \mathbf{B}_M)$, satisfying the consistency condition, be determined from finite sets $M \subset T$. Then there exists a unique probability measure P on $(X_T, \mathbf{B}_T)$, whose projection onto $(X_M, \mathbf{B}_M)$ coincides with P_M for any set $M \subset T$, i.e.,

$$P(A \times X_{T \setminus M}) = P_M(A) \text{ for all } A \in \mathbf{B}_M.$$

Corollary. Assume that the conditions of the theorem are fulfilled. Then there exists a random process $\xi = \{\xi(t) : t \in T\}$ such that

(1) the RV $\xi(t)$ assumes values in X_t for any $t \in T$, and,

(2) for any finite set $M = \{t_1, \ldots, t_n\} \subset T$,

$$P\{\xi(t_1) \in B_1, \ldots, \xi(t_n) \in B_n\} = P_M(B)$$

for all

$$B = B_1 \times \ldots \times B_n \in \mathbf{B}_M, \text{ where } B_k \in \mathbf{B}_{t_k}$$

(i.e., the finite-dimensional distributions of the process ξ coincide with the corresponding probability measures of the given consistent family of measures).

In fact, it suffices to put

$$\xi(t, \omega) = \omega(t) \text{ for } t \in T, \ \omega \in X_T,$$

since

$$\omega = \{\omega(t) : t \in T\} \in X_T,$$

and

$$\begin{aligned} &P\{\xi(t_1) \in B_1, \ldots, \xi(t_n) \in B_n\} \\ &= P\{\omega : \xi(t_k, \omega) \in B_k, \ k = 1, \ldots, n\} \\ &= P\{\omega : \omega(t_k) \in B_k, \ k = 1, \ldots, n\} \\ &= P\{\omega \in X_T : \omega \in B \times X_{T \times M}\} \\ &= P(B \times X_{T \times M}) = P_M(B). \end{aligned}$$

Remark. Similarly to the definition of the measurable space $(X_T, \mathbf{B}_T)$ and (by the theorem) of the measure $P_T = P$ on this space, proceeding from the family of Polish spaces (X_t, d_t), $t \in T$, and the consistent family of probability measures P_M, $M \subset T$, $|M| < \infty$, we can define the measurable space $(X_S, \mathbf{B}_S)$ and measure P_S on this space for any set $S \subset T$, (not necessarily finite), proceeding

from the family of Polish spaces (X_t, d_t), $t \in T$, and the consistent family of probability measures P_M, $M \subset S$, $|M| < \infty$. Meanwhile, the projection of the measure P_T onto $(X_S, \mathbf{B}_S)$ coincides with P_S, i.e., $P_T(A \times X_{T \setminus S}) = P_S(A)$ for all $A \in \mathbf{B}_S$. Indeed, put

$$Q_S(A) = P_T(A \times X_{T \setminus S})$$

for all $A \in \mathbf{B}_S$. It should be shown that $P_S = Q_S$. But this follows from the projections of measures P_S and Q_S onto $(X_M, \mathbf{B}_M)$ coinciding with P_M for any finite set $M \subset S$, and uniqueness (in view of the theorem) of a probability measure on $(X_S, \mathbf{B}_S)$ with such a property.

Example 3. Let F be a DF on R^1. We show that there exists a sequence $\{\xi_n : n = 1, 2, \ldots\}$ of i.i.d. RRV's with the common DF F. Put $T = \{1, 2, \ldots\}$, $X_t = R^1$ for $t \in T$, and define a probability measure P_M on $(X_M, \mathbf{B}_M) = (R^n, \mathbf{B}^n)$ for $M \subset T$ and $|M| = n$ by putting

$$P_M(B) = \prod_{t \in M} Q(B_t)$$

for a rectangle $B = B_1 \times \ldots \times B_n \in \mathbf{B}^n$, where $B_k \in \mathbf{B}^1$ and Q is a probability measure on R^1, corresponding to the DF F. Since the set of rectangles forms a semi-algebra, and P_M is probability on this semi-algebra, it is uniquely extendable to the minimal σ-algebra containing these rectangles, i.e., to $\mathbf{B}^n$. The family of probability measures P_M with respect to finite sets $M \subset T$ satisfies the consistency condition, since, for $N \subset M \subset T$ and rectangle $B = \prod_{t \in N} B_t \in \mathbf{B}_N$, we have

$$P_M(B \times X_{M \setminus N}) = \prod_{t \in N} Q(B_t) \cdot \prod_{t \in M \setminus N} Q(X_t) = \prod_{t \in N} Q(B_t) = P_N(B)$$

due to $Q(X_t) = Q(R^1) = 1$. By the corollary to the theorem, there exists a process $\{\xi_t : t \in T\} = \{\xi_n : n = 1, 2, \ldots\}$ satisfying the condition

$$P\{\xi_1 \in B_1, \ldots, \xi_n \in B_n\} = P_{\{1, \ldots, n\}}(B_1 \times \ldots \times B_n) = \prod_{k=1}^{n} Q(B_k)$$

for any integer $n \geqslant 1$ and $B_k \in \mathbf{B}^1$ when $k = 1, \ldots, n$. Selecting $B_i = R^1$ for $i \neq k$, we obtain

$$P\{\xi_k \in B_k\} = P\{\xi_1 \in B_1, \ldots, \xi_n \in B_n\} = \prod_{i=1}^{n} Q(B_i)$$
$$= Q(B_k) \cdot \prod_{i \neq k} Q(B_i) = Q(B_k).$$

Therefore,

$$P\{\xi_1 \in B_1, \ldots, \xi_n \in B_n\} = \prod_{k=1}^{n} P\{\xi_k \in B_k\}$$

for all $n \geqslant 1$ and $B_k \in \mathbf{B}^1$, i.e., the RV's $\{\xi_n : n \geqslant 1\}$ are independent, and each has the distribution function

$$P\{\xi_k < x\} = P\{\xi_k \in B = (-\infty, x)\} = Q(B) = F(x).$$

Proof of the theorem. We shall first verify **P*1**: the set $\mathbf{C}$ of all cylinders is an algebra of sets.

Then we define a measure P on $\mathbf{C}$ by putting $P(C) = P_M(A)$ for the cylinder $C = A \times X_{T \setminus M}$, and verify

P2: the definition of P is correct,

P3: P is additive,

and

P4: P is continuous in $\varnothing$, i.e., it follows from $\mathbf{C} \ni C_n \downarrow \varnothing$ that $P(C_n) \to 0$.

Since a measure P is normalized, i.e., $P(X_T) = 1$ for $X_T \in \mathbf{C}$, it follows that P is probability on $\mathbf{C}$ (normed, nonnegative and σ-additive measure), and, therefore, admits a unique extension to $\mathbf{B}_T$, the minimal σ-algebra containing $\mathbf{C}$.

Thus, it now suffices to verify the validity of P1-P4.

P1. It follows from the equality $(A \times X_{T \setminus M})^c = A^c \times X_{T \setminus M}$ that the complement to a cylinder is a cylinder. We shall then see that the intersection of cylinders is a cylinder.

Let

$$C_i = A_i \times X_{T \setminus M_i} \in \mathbf{C}, \quad i = 1, 2.$$

Put

$$M = M_1 \cup M_2,$$
$$B_1 = A_1 \times X_{M \setminus M_1} \subset X_M, \qquad B_2 = A_2 \times X_{M \setminus M_2} \subset X_M,$$
$$A = B_1 \cap B_2 \subset X_M, \quad C = A \times X_{T \setminus M}.$$

Then $C_1 \cap C_2 = C$, and C is a cylinder.

P2. It should be shown that $P(C)$ does not depend on the representation of the cylinder C in the form $C = A \times X_{T \setminus M}$. Let

$$C = A_1 \times X_{T \setminus M_1} = A_2 \times X_{T \setminus M_2}.$$

We shall show that

$$P_{M_1}(A_1) = P_{M_2}(A_2).$$

Making use of the notation from P1 and the consistency condition, we obtain

$$P_M(B_1) = P_{M_1}(A_1), \qquad P_M(B_2) = P_{M_2}(A_2).$$

It remains to notice that the equality

$$B_1 \times X_{T \setminus M} = A_1 \times X_{T \setminus M_1} = A_2 \times X_{T \setminus M_2} = B_2 \times X_{T \setminus M}$$

yields $B_1 = B_2$.

* P stands for *Proposition*.

P3. Let two cylinders C_1 and C_2 be non-intersecting. Using the notation of P1, and also the statement of P2, we have

$$C_1 = B_1 \times X_{T \setminus M}, \quad C_2 = B_2 \times X_{T \setminus M},$$

$$P(C_1 + C_2) = P\{(B_1 + B_2) \times X_{T \setminus M}\} = P_M(B_1 + B_2)$$
$$= P_M(B_1) + P_M(B_2) = P(C_1) + P(C_2).$$

P4. Let $C_n = A_n \times X_{T \setminus M_n}$, $|M_n| < \infty$, $n \geqslant 1$. We can assume that $M_1 \subset M_2 \subset \ldots$. Suppose that $P(C_n) \geqslant \varepsilon_0 > 0 \ \forall n \geqslant 1$, and show that this contradicts $C_n \downarrow \varnothing$, which means that $C_n \downarrow \varnothing$ implies $P(C_n) \to 0$.

1. Let us first prove that we can restrict ourselves to the case where A_n is a compactum in X_{M_n}, $n \geqslant 1$, for which we shall need the following statement.

Lemma. Let P be a probability measure on a Polish space. For any $\varepsilon > 0$ and any Borel set A in this space, there exists a compactum K such that

$$K \subset A, \text{ and } P(A \setminus K) \leqslant \varepsilon.$$

We also note that the direct product of a finite number of Polish spaces is a Polish space.

In our case, for any $n \geqslant 1$, we select (according to the lemma) a compactum K_n in X_{M_n}, so that

$$K_n \subset A_n, \text{ and } P_{M_n}(A_n \setminus K_n) = P(C_k \setminus \mathscr{K}_n) \leqslant \frac{\varepsilon_0}{2^{n+1}}$$

where $\mathscr{K}_n$ is a cylinder with base K_n. Putting

$$C_n^* = \bigcap_{i=1}^{n} \mathscr{K}_i,$$

we have

$$P(C_n \setminus C_n^*) = P \bigcup_{i=1}^{n} (C_n \setminus \mathscr{K}_i) \leqslant P \bigcup_{i=1}^{n} (C_i \setminus \mathscr{K}_i)$$
$$\leqslant \sum_{i=1}^{n} P(C_i \setminus \mathscr{K}_i) \leqslant \frac{\varepsilon_0}{2}$$

whence

$$\varepsilon_0 \leqslant P(C_n) \leqslant P(C_n^*) + \varepsilon_0/2, \quad \text{i.e., } P(C_n^*) \geqslant \varepsilon_0/2 > 0.$$

Thus, $P(C_n^*) \geqslant \varepsilon_0^* = \varepsilon_0/2 \ \forall n \geqslant 1$, and the compactum $A_n^* = \bigcap_{i=1}^{n} K_i \times X_{M_n \setminus M_i} \subset K_n$ in X_{M_n} serves as the base of the cylinder C_n^*. Besides,

$$C_n^* \subset C_n, \quad A_n^* \subset A_n, \quad C_1^* \supset C_2^* \supset \ldots.$$

2. Thus, let $P(C_n) \geqslant \varepsilon_0 > 0$, and A_n a compactum in X_{M_n}, $n \geqslant 1$. It remains to show that this contradicts $C_n \downarrow \varnothing$. We select

$x^{(n)} = \{x^{(n)}(t) : t \in T\}$ in each cylinder C_n. This can always be done, since $C_n = \varnothing$ (because $P(C_n) > 0$). Besides, since $C_n = A_n \times X_{S \setminus M_n} \times X_{T \setminus S}$, where $S = \bigcup_{n \geqslant 1} M_n = \lim M_n$, we can assume that

$$x^{(n)}(t) = x^{(1)}(t) = w(t) \text{ for } t \in T \setminus S.$$

Denote the restriction of the function $x^{(n)}$ from T to M by $x_M^{(n)} = \{x^{(n)}(t) : t \in M\}$. Since the points of $x_{M_1}^{(n)}$ belong to the compactum $A_1 \subset X_{M_1}$, there exists a subsequence $\{n_1\} \subset \{n\}$ such that the sequence $\{x_{M_1}^{(n_1)}\} \subset A_1$ converges to a certain point $w_{M_1} = \{w(t) : t \in M_1\} \subset A_1$. Similarly, since the points of $x_{M_2}^{(n_1)}$ belong to the compactum $A_2 \subset X_{M_2}$, there exists a subsequence $\{n_2\} \subset \{n_1\}$ such that the sequence $\{x_{M_2}^{(n_2)}\} \subset A_2$ converges to a certain point $w_{M_2} = \{w(t) : t \in M_2\} \in A_2$. Subsequences $\{n\} \supset \{n_1\} \supset \{n_2\} \supset \ldots \supset \{n_k\} \supset \ldots$ are constructed in the same way, so that the sequence $\left\{x_{M_k}^{(n_k)}\right\}$ of points from the compactum $A_k \subset X_{M_k}$ converges to a certain point $w_{M_k} = \{w(t) : t \in M_k\} \in A_k$. Selecting a diagonal sequence $\{d_k\} \subset \{n\}$ (where d_k is the k-th term in the sequence $\{n_k\}$), we obtain that, for a certain point $w = \{w(t) : t \in T\}$,

$$x^{(d_k)}(t) \to w(t) \ \forall t \in T$$

($x^{(d_k)}(t) = w(t)$ when $t \in T \setminus S$), and $w_{M_k} \in A_k$, $k \geqslant 1$, which means that $w \in C_n$ for all $n \geqslant 1$; therefore, $\bigcap_{n \geqslant 1} C_n \neq \varnothing$.

Chapter 3

Discrete- and Continuous-Time Markov Chains

§ 1. Definition of Discrete- and Continuous-Time Markov Chains

A process $\xi = \{\xi(t) : t \in T\}$, where the set T is at most countable, is called a **discrete-time process** (or **random sequence**). As a rule, the set T coincides with the set of integers or its part in this case.

If, however, T coincides with a straight line interval, e.g., $T = (-\infty, \infty)$, or $(0, \infty)$, or $[a, b]$, or (a, b), etc., then the random process ξ is called a **continuous-time process.**

1°. Consider a sequence $\{\xi_n\}$ of RV's which, according to new terminology, is a discrete-time random process. Suppose that each RV ξ_n assumes values from the set $E = \{0, 1, 2, \ldots\}$, and con-

sider the following simple form of dependence of RV's from the sequence $\{\xi_n\}$, viz., the distribution of an RV ξ_{n+1} may depend on the value assumed by the RV ξ_n, and does not depend on the values taken by the preceding RV's. E.g., $\xi_{n+1} = h_n(\xi_n, \eta_n)$, $n = 1, 2, \ldots$, where the (Borel) functions h_n are given, the RV's $\eta_1, \eta_2, \ldots$ are independent of each other, and do not depend on the RV's $\xi_1, \xi_2, \ldots$, respectively (i.e., η_n does not depend on ξ_n). Figuratively speaking, if the "present" time is fixed, then the "future" time does not depend on the "past". Formally, this means that

$$P(\xi_{n+1} = i_{n+1} \mid \xi_1 = i_1, \ldots, \xi_n = i_n) = P(\xi_{n+1} = i_{n+1} \mid \xi_n = i_n) \tag{1}$$

for all $n \geqslant 1$ and $i_k \in E$, $1 \leqslant k \leqslant n+1$, such that

$$P(\xi_1 = i_1, \ldots, \xi_n = i_n) > 0.$$

The RV's $\{\xi_n\}$ satisfying (1) are said to **form a Markov chain**, and the sequence $\{\xi_n\}$ itself is called a **discrete-time Markov chain**. Such a chain is said to be **homogeneous** (with respect to time) if, for all $i, j \in E$, probability

$$P(\xi_{n+1} = j \mid \xi_n = i) = p_{ij}$$

does not depend on the "time" n. The **evolution** of a homogeneous Markov chain $\{\xi_n\}$ is determined by

(a) the initial distribution of the RV ξ_1, i.e., set of numbers

$$P(\xi_1 = i) = p_i, \quad i \in E, \ p_i \geqslant 0, \quad \Sigma p_i = 1,$$

and

(b) the **one-step transition probability matrix** $\{p_{ij}\}$.

Meanwhile, the set E is called the **state space (phase space)** of the chain, and probability

$$P(\xi_{n+1} = j \mid \xi_i = i) = p_{ij}^n$$

n-step probability of transition from a state $i \in E$ to a state $j \in E$. In particular, $p_{ij}^1 = p_{ij}$. It follows from (1) that

$$p_{ij}^{m+n} = \sum_{k \in E} p_{ik}^m p_{kj}^n. \tag{2}$$

Remark 1. Usually, the **minimal** state space is taken as the phase space E, which means that, for any $j \in E$, there exist $i \in E$ and "time" $n \geqslant 1$ such that $p_{ij}^n > 0$ (otherwise, it would mean that there is a state j into which we cannot go in any time n, starting with any i).

2°. Consider a random process $\xi = \{\xi(t) : t \in T = (0, \infty)\}$ taking values from $E = \{0, 1, 2, \ldots\}$, and assume that

$$P\{\xi(t_{n+1}) = i_{n+1} \mid \xi(t_1) = i_1, \ldots, \xi(t_n) = i_n\} = P\{\xi(t_{n+1}) = i_{n+1} \mid \xi(t_n) = i_n\} \tag{3}$$

for any integer $n \geqslant 1$, any moments $t_1 < \ldots < t_{n+1}$ from T, and any states $i_1, \ldots, i_{n+1}$ from E, so that

$$P\{\xi(t_1) = i_1, \ldots, \xi(t_n) = i_n\} > 0.$$

The RV's $\{\xi(t) : t \in T\}$ satisfying (3) are said to form a Markov chain, and the random process ξ itself is called a **continuous-time Markov chain**. Such a chain is said to be homogeneous (with respect to time) if

$$P\{\xi(s+t) = j \mid \xi(s) = i\} = p_{ij}^t$$

does not depend on $s \in T$ for any i, j from E and $t \in T$. Meanwhile, p_{ij}^t is called probability of transition from a state i to a state j in time t. It follows from (3) that

$$p_{ij}^{s+t} = \sum_k p_{ik}^s p_{kj}^t. \tag{4}$$

3°. **Definition.** A random process $\{\xi(t) : t \in T\}$, where $T = \{1, 2, \ldots\}$ or $T = (0, \infty)$, assuming values from $E = \{0, 1, 2, \ldots\}$ will be called a homogeneous Markov chain (in the wide sense) if probabilities

$$P\{\xi(s+t) = j \mid \xi(s) = i\} = p_{ij}^t$$

do not depend on $s \in T$ for any $i, j \in E$ and $t \in T$, satisfy the Chapman-Kolmogorov equation

$$p_{ij}^{s+t} = \sum_k p_{jk}^s p_{kj}^t$$

for all $s, t \in T$, $i, j \in E$, and

$$\sum_j p_{ij}^t = 1$$

for all $i \in E$ and $t \in T$.

Remark 2. In this definition, we do not assume conditions (1) or (3) to be fulfilled.

Remark 3. A homogeneous Markov chain determines a family of matrices $\{P^t : t \in T\}$ satisfying the conditions

(a) for each $t \in T$, the matrix $P^t = \{p_{ij}^t\}$ is **stochastic**, i.e.,

$$p_{ij}^t \geqslant 0 \text{ and } \sum_{j \geqslant 0} p_{ij}^t = 1 \ \forall i, j, t,$$

and

(b) $$P^{s+t} = P^s \cdot P^t \quad \forall s, t \in T.$$

Condition (b) is just another form of the Chapman-Kolmogorov equation. Conversely, any family of matrices $\{P^t : t \in T\}$ satisfying (a) and (b) determines a certain homogeneous Markov chain $\{\xi(t) : t \in T\}$ with the matrix of transition probabilities P^t (to prove this fact, make use of the Kolmogorov theorem; see the introduction to Part 2).

Remark 4. As regards Markov chains, the following problems are usually posed.

(1) Find the existence conditions for $\lim_{t\to\infty} p_{ij}^t$, and indicate methods of determining these limits.

(2) Clarify the structure of a homogeneous Markov chain.

(3) Determine how the laws of large numbers (ergodic theorems) are related to the sojourn time in a given state space.

(4) Formulate the Central Limit Theorems for a sequence of RV's forming a Markov chain.

(5) Clarify the asymptotic behaviour of functionals of Markov chains, different from (3) and (4).

In this chapter, we only investigate problems (1)-(3).

§ 2. Examples of Homogeneous Markov Chains

Example 1. Consider a random walk of a particle on integral points in an interval $[a, b]$, where a and b are integers, $a < b$. The particle at a point (state) i, $a \leqslant i \leqslant b$, goes to the neighbouring state $i+1$

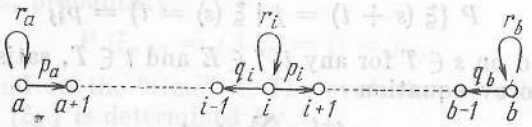

Fig. 7

or $i-1$ in one step or remains in the state i with probabilities p_i, q_i and r_i, respectively ($p_i + q_i + r_i = 1$). Certainly, $q_a = 0 = p_b$ (see Fig. 7).

Let the particle start from a certain state $\xi_0 \in [a, b]$. Denote its state in n steps by ξ_n. The RV's of the sequence $\{\xi_n\}$ form a homogeneous Markov chain with transition probabilities

$$p_{ii+1} = p_i,\ p_{ii-1} = q_i,\ p_{ii} = r_i,$$
$$q_a = 0 = p_b,\ p_{ij} = 0 \text{ for } |i-j| > 1.$$

We now list certain special cases where it is assumed that $p_i > 0$ and $q_i > 0$ for $a < i < b$.

Case 1: $r_a = 1$. This means that, on reaching the state a, the particle remains in this state forever, due to which the state a is often called an **absorbing barrier**. Similarly, if $r_b = 1$, then the state b is for the same reason called an absorbing barrier (on reaching the state b, the particle remains in this state forever).

Case 2: $a = -\infty$, $b = +\infty$, $r_i = 0$, $p_i = p\ \forall i$. Under these conditions, the walk of the particle is often called a **free random walk** on integral points of the straight line. In particular, if $p_i = p = \frac{1}{2}\ \forall i$, then the walk is said to be **symmetric.**

Case 3: $r_i = 0$, $p_i = p$ for $a < i < b$, $p_a = p$, $r_a = 1-p$, $r_b = p$, $q_b = 1-p$ (see Fig. 8). Such a walk of the particle in $[a, b]$ is obtained from a free random walk if each time it gets into the state $a-1$, it then gets into the state a. If, however, the particle gets into the state $b+1$, then it is carried into the state b. This walk is sometimes interpreted as the process of displacing the particle with two "reflecting barriers" situated at the points $a - \frac{1}{2}$ and $b + \frac{1}{2}$. Due to such an interpretation, this case has the name of a **random**

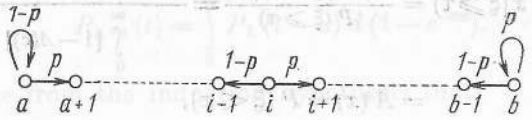

Fig. 8

walk of a particle **with two reflecting barriers.** If $b = +\infty$, then we obtain a **random walk of the particle with one reflecting barrier.**

Example 2 (*Poisson process*). Let $\{z_k\}$ be a sequence of independent positive semi-definite RV's having the same exponential distribution

$$P(z_k < x) = 1 - e^{-ax}, \quad x \geqslant 0,\ k \geqslant 1,$$

with parameter a. Put $t_0 = 0$, $t_k = z_1 + \ldots + z_k$, $k \geqslant 1$, $t_{k+1} \geqslant t_k$. Denote the number of points of the sequence $\{t_k\}$, which are to the left of a point t, by

$$\nu(t) = \sup\{k:\ t_k < t\}.$$

E.g., if $z_1, z_2, \ldots$ are the lengths of life of the sequence of replaceable elements, then $\nu(t)$ is the number of elements replaced up to the epoch t. The process $\nu(t)$, $t > 0$, is a homogeneous Markov chain (in the wide sense). To prove this statement, it suffices to verify the formula

$$(1) \qquad P\{\nu(s+t) = j / \nu(s) = i\} = p_{ij}^t = \frac{(at)^{j-i}}{(j-i)!} e^{-at}$$

for $j \geqslant i$, and $p_{ij}^t = 0$ for the remaining i and j (thereby, p_{ij}^t does not depend on s), since the Chapman-Kolmogorov equation

$$p_{ij}^{s+t} = \sum_k p_{ik}^s p_{kj}^t$$

follows from this formula automatically.

To prove (1), we make use of the following statement.

Lemma. Let ξ and τ be two independent positive semi-definite RV's, and the RV ξ obey the exponential distribution. Then

$$P(\xi - \tau < x \mid \xi \geqslant \tau) = P(\xi < x)\ \forall x.$$

Remark 1. In particular, this lemma is valid for any nonnegative member τ. E.g., if the duration of a telephone talk obeys the exponential distribution, then the remaining time interval does not depend on the time of talk elapsed, and obeys the same exponential distribution.

Proof of the lemma may be sufficient for $x > 0$. If A and F are the DF's of two RV's ξ and τ, respectively, then

$$P(\xi - \tau < x | \xi \geqslant \tau) = \frac{P(\tau \leqslant \xi < x+\tau)}{P(\xi \geqslant \tau)} = \frac{\int_0^\infty [A(x+c) - A(c)]\, dF(c)}{\int_0^\infty [1 - A(c)]\, dF(c)}$$

$$= A(x) = P(\xi < x),$$

where we have taken into account that

$$A(x+c) - A(c) = A(x)[1 - A(c)],$$

since the DF A is of the form

$$A(x) = 1 - e^{-ax}, \quad x \geqslant 0, \quad a > 0.$$

We now show that

$$P\{\nu(s+t) = i + k / \nu(s) = i\} = P\{\nu(t) = k\} \tag{2}$$

for all $s > 0$, $t > 0$, $i \geqslant 0$, $k \geqslant 0$. It is clear that

$$P\{\nu(t) = k\} = P\{t_k < t,\ t_{k+1} \geqslant t\}. \tag{3}$$

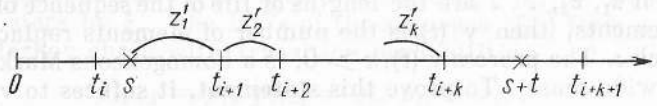

Fig. 9

Put (see Fig. 9)

$$z_1' = z_{i+1} - (s - t_i), \qquad z_2' = z_{i+2}, \qquad z_3' = z_{i+3}, \quad \ldots,$$

$$t_0' = 0, \qquad t_k' = z_1' + \ldots + z_k', \ k \geqslant 1, \qquad t_{i+k} = s + t_k'.$$

We obtain

$$P\{\nu(s+t) = i + k | \nu(s) = i\} = P(t_k' < t,\ t_{k+1}' \geqslant t | \nu(s) = i). \tag{4}$$

By the lemma, the RV $z_1' = z_{i+1} - \tau$, where $\tau = s - t_i$, has the same exponential distribution as the RV z_{i+1} if $\tau > 0$ and $z_1' \geqslant 0$,

i.e., if $t_i < s$ and $t_{i+1} \geqslant s$, or if $\nu(s) = i$. Thus, provided $\{\nu(s) = i\}$, the RV's $\{z_k',\ k \geqslant 1\}$ are i.i.d. exponentially with the same parameter a. But then (2) follows from (3)-(4).

It now remains to prove the formula

$$P_k(t) = P\{\nu(t) = k\} = \frac{(at)^k}{k!} e^{-at}, \quad k \geqslant 0, \quad t \geqslant 0. \tag{5}$$

For $k = 0$, we have $P_0(t) = P(t_1 \geqslant t) = e^{-at}$, and relation (5) is valid. Let formula (5) hold for a certain $k \geqslant 0$. It follows from the definition of $\nu(t)$, $t \geqslant 0$, that

$$P_{k+1}(t) = \int_0^t P_k(t-x)\, d(1 - e^{-ax}).$$

We derive from the induction hypothesis that

$$P_{k+1}(t) = \int_0^t \frac{[a(t-x)]^k}{k!} e^{-a(t-x)} \cdot a e^{-ax}\, dx = e^{-at} \frac{(at)^{k+1}}{(k+1)!},$$

which coincides with (5) if k is replaced by $k + 1$. Formula (5) is thus proved.

Hence, the RV $\nu(t)$ has the Poisson distribution with parameter at for any $t \geqslant 0$. The process $\{\nu(t) : t \geqslant 0\}$ is said to be **Poisson.**

Example 3. Consider a homogeneous discrete-time Markov chain $\{\xi_n\}$, determined by a one-step transition probability matrix $P = \{p_{ij}\}$. Note that if $P^{(k)} = \{p_{ij}^{(k)}\}$ is a k-step transition probability matrix, then, in view of the Chapman-Kolmogorov equation,

$$P^{(k)} = P^k, \ k = 1, 2, \ldots.$$

We shall now assume that each step lasts for a random time, the durations of the steps are independent, and have the exponential distribution with parameter a. In other words, a change of state may occur at the epochs of jumps in a Poisson process (see the previous example), and is controlled by the transition probability matrix P. Denote the state at epoch t by $\xi(t)$; in other words, $\xi(t) = \xi_{\nu(t)}$, where $\nu(t)$ is the number of jumps in the Poisson process up to epoch t. Then

$$p_{ij}^t = P\{\xi(s+t) = j \mid \xi(s) = i\} = \sum_{k \geqslant 0} \frac{(at)^k}{k!} e^{-at} p_{ij}^{(k)},$$

where $p_{ij}^{(0)} = \delta_{ij}$. The Chapman-Kolmogorov equation is valid for transition probabilities $\{p_{ij}^{(t)}\}$ if it is valid for $\{p_{ij}^{(k)}\}$. Putting $P(t) = \{p_{ij}^t\}$, we obtain

$$P(t) = \sum_{k \geqslant 0} \frac{(at)^k}{k!} e^{-at} P^k,$$

i.e.,

$$P(t) = e^{at(P-I)};$$

where $I = \{\delta_{ij}\}$. Thus, $\{\xi(t) : t \geqslant 0\}$ is a homogeneous continuous-time Markov chain. Actually, each homogeneous continuous-time Markov chain with a finite number of states is arranged in the same way.

Example 4. Consider a walk of a particle on the points of a set $E \subset \{0, 1, 2, \ldots\}$. Some states (points) from E are regarded as absorbing. Let the particle change its state into a state i at a certain epoch. If i is an absorbing state, then the particle remains in it forever. Otherwise, the sojourn time at the state i is determined by the exponential distribution $1 - \exp\{-\alpha_i x\}$, $\alpha_i > 0$, $x \geqslant 0$, and does not depend on the previous trajectory of the particle. The particle then changes over to a state j with probability q_{ij}, etc.; $q_{ii} = 0$. Let $\xi(t)$ be the state of the particle at an epoch t, beginning with a certain state $\xi(0) \in E$. A homogeneous continuous-time Markov chain $\{\xi(t),\ t \geqslant 0\}$ is called a **minimal Markov chain (process).** It turns out that, under quite general assumptions, a homogeneous continuous-time Markov chain is equivalent (in the sense of coincidence of finite-dimensional distributions) to a minimal Markov chain.

Remark 2. The definition of a minimal Markov chain can be made more correct by the following procedure. Consider a sequence

$$\zeta_n = (z_n, \xi_n), \quad n = 0, 1, 2, \ldots$$

of RV's with values in $[0, \infty) \times E$, forming a homogeneous Markov chain with transition probabilities of the form

$$P(z_{n+1} < y, \xi_{n+1} = j \mid z_n = x, \xi_n = i) = P(z_{n+1} < y, \xi_{n+1} = j \mid \xi_n = i) = q_{ij}(1 - e^{-\alpha_i y}).$$

Specifying the initial distribution of the RV ξ_0, we establish the existence of such a chain by the Kolmogorov theorem. Assuming $z_0 = 0$, we introduce the following interpretation of the chain. A certain particle is at the state ξ_0 in a time interval of length z_1, at the state ξ_1 in the next interval of length z_2, etc., i.e., in the interval $t_n \leqslant t < t_{n+1}$, where $t_0 = 0$, $t_n = z_1 + \ldots + z_n$, it is at the state ξ_n. It remains to put

$$\xi(t) = \xi_n \text{ if } t_n \leqslant t < t_{n+1}.$$

In other words, $\xi(t) = \xi_{\nu(t)}$, where $\nu(t) = \sup\{n : t_n \leqslant t\}$. It follows (show!) from the construction and fundamental property of the exponential distribution (see Lemma) that $\{\xi(t),\ t \geqslant 0\}$ is a homogeneous continuous-time Markov chain.

Example 5 (*birth-and-death process*). Consider a minimal continuous-time Markov process with the state space $E = \{0, 1, 2, \ldots, N\}$ or $E = \{0, 1, 2, \ldots\}$ again, but now assuming that jumps may occur only into the neighbouring states, which means that $q_{ij} = 0$ when $|i - j| > 1$. Put $q_{ii+1} = p_i$ when $i \geqslant 0$, and $q_{ii-1} = q_i = 1 - p_i$ when $i \geqslant 1$. Such a process (starting with a certain state $\xi(0) \in E$) is called a **birth-and-death process.**

If $p_i = 1$ for all i, then a **pure birth process** is spoken of. E.g., such is a Poisson process (see Example 2). If, however, $p_i = 0$ for all i, then the process is called a **pure death process.**

Often, when describing dynamic systems (systems developing with time) by a birth-and-death process, the following simple properties of exponentially distributed RV's may prove useful (they complete the basic property of the exponential distribution; see Lemma). Let $\xi_1, \ldots, \xi_n$ be independent RV's having the exponential distribution with parameters $a_1, \ldots, a_n$, respectively. Then

(1) the RV $\min(\xi_1, \ldots, \xi_n)$ has the exponential distribution with parameter $a = a_1 + \ldots + a_n$, and

$$(2) \qquad P(\xi_1 < \xi_2) = \frac{a_1}{a_1 + a_2}.$$

§ 3. Limiting Distribution of Markov Chains

Consider a homogeneous Markov chain determined by the matrix of transition probabilities $P^t = \{p_{ij}^t\}$, where $T = \{1, 2, \ldots\}$ or $T = (0, \infty)$, and $E = \{0, 1, 2, \ldots\}$ is the state space of the chain.

We call a Markov chain **contracting** if, for any pair of states $i_1, i_2 \in E$, there exist a time $t \in T$ and state $j \in E$ such that $p_{i_1 j}^t > 0$, and $p_{i_2 j}^t > 0$ (i.e., the system can change from any pair of states into another during the same time).

A distribution $\pi = \{\pi_i\}$ on E (i.e., a set of numbers $\pi = (\pi_0, \pi_1, \pi_2, \ldots)$ such that $\pi_i \geqslant 0$, and $\sum_i \pi_i = 1$) is said to be **stationary** if

$$\sum_i \pi_i p_{ij}^t = \pi_j \text{ for all } t \in T,\ j \in E.$$

Theorem 0. If a homogeneous discrete- or continuous-time Markov chain assumes a finite number of states, then the following statements are equivalent.

1°. There exists a unique stationary distribution $\pi = \{\pi_i\}$, and $p_{ij}^t \to \pi_j$ at $t \to \infty$ for any i, j.

2°. The chain is contracting.

The following theorems touch upon the general case where the chain does not necessarily assume a finite number of values.

Theorem 1. Let a homogeneous Markov chain be contracting. Then, for any two states i and j from E, there exists $\lim_{t \to \infty} p_{ij}^t$. Meanwhile,

1°. If there exists no stationary distribution, then $p_{ij}^t \to 0$ for all i, j from E (as $t \to \infty$).

2°. If there exists a stationary distribution $\pi = \{\pi_i\}$, then it is unique, and $p_{ij}^t \to \lambda_i \pi_j$ as $t \to \infty$ for all i and i, where λ_i possibly depends on i, while $0 < \lambda_i \leqslant 1$.

Theorem 2. Let there exist a stationary distribution $\pi = \{\pi_i\}$. Then the following statements are equivalent.

1°. $\lim_{t\to\infty} p_{ij}^t = \pi_j \ \forall i,\ j.$

2°. The chain is contracting, and each bounded solution of the system of equations

$$\sum_j p_{ij}^t x_j = x_i \ \forall i,\ t$$

is of the form

$$x = (x_0,\ x_1,\ \ldots) = c\ (1,\ 1,\ \ldots).$$

A Markov chain is said to be **irreducible** if, for any pair of states i and j, there exists $t \in T$ such that $p_{ij}^t > 0$.

Theorem 3. Let there exist a stationary distribution $\pi = \{\pi_i\}$. Then the following statements are equivalent.

1°. $\lim_{t\to\infty} p_{ij}^t = \pi_j > 0 \ \forall i,\ j.$

2°. The chain is contracting and irreducible.

3°. For any pair of states i and j, there exists a time $\tau = \tau\ (i, j) \in T$ such that $p_{ij}^t > 0$ for all $t \geqslant \tau$.

If $T = \{1,\ 2,\ \ldots\}$, then each of these statements is equivalent to the following.

4°. The chain is aperiodic and irreducible.

Meanwhile, a homogeneous Markov chain is said to be **aperiodic** if, for any pair of states i and j, the greatest common divisor of the numbers $\{n : p_{ij}^n > 0\}$ equals unity.

Remark. A distribution $\alpha = \{\alpha_i\}$ on E (i.e., a set of numbers $\alpha = \{\alpha_i\}$ such that $\alpha_i \geqslant 0$, and $\sum_i \alpha_i = 1$) is called the **limiting distribution** of the Markov chain if, for a certain state $i \in E$,

$$\lim_{t\to\infty} p_{ij}^t = \alpha_j \ \forall j.$$

Note that a limiting distribution is stationary, i.e., satisfies the system of equations

$$\sum_i \alpha_i p_{ij}^t = \alpha_j \ \forall j,\ t.$$

In fact, if the number of states is finite, we obtain from the Chapman-Kolmogorov equation

$$p_{ij}^{s+t} = \sum_k p_{ik}^s p_{kj}^t$$

as $s \to \infty$ that

$$\alpha_j = \sum_k \alpha_k p_{kj}^t,$$

i.e., $\alpha = \{\alpha_i\}$ is a stationary distribution. In the general case, we can use the fact that $p_{ij}^t \to \alpha_j$ uniformly (with respect to j). Indeed,

for $\varepsilon > 0$, we select N such that

$$\sum_{k\geqslant N} \alpha_k < \varepsilon,$$

and then such s_k, $0 \leqslant k < N$, that

$$|\ p_{ik}^s - \alpha_k\ | < \varepsilon/N \text{ when } s \geqslant s_k, \text{ and } k < N.$$

Hence, when $j \geqslant N$,

$$|\ p_{ij}^s - \alpha_j\ | \leqslant p_{ij}^s + \alpha_j \leqslant \sum_{k\geqslant N} p_{ik}^s + \sum_{k\geqslant N} \alpha_k \leqslant 1 - \sum_{k<N} p_{ik}^s + \varepsilon$$
$$= \left(1 - \sum_{k<N} \alpha_k\right) + \varepsilon - \sum_{k<N} (p_{ik}^s - \alpha_k)$$
$$\leqslant 2\varepsilon + \sum_{k<N} |\ p_{ik}^s - \alpha_k\ | < 3\varepsilon$$

for $s \geqslant \bar{s} = \max\ (s_0,\ s_1,\ \ldots,\ s_{N-1})$. Thus,

$$|\ p_{ij}^s - \alpha_j\ | < 3\varepsilon \ \forall j, \quad s \geqslant \bar{s}.$$

§ 4. Proofs

A. Theorem 0 is a simple corollary to Theorems 1-2. In fact, $1° \Rightarrow 2°$ by Theorem 2, and $2° \Rightarrow 1°$ by Theorem 1, since, in the case of a finite number of states, $\alpha = \{\alpha_j = \lim_{t\to\infty} p_{ij}^t\}$ is a limiting distribution, and, therefore, a stationary distribution according to the Remark in § 3.

To prove Theorems 1-3, we make use of Propositions 1-3 from functional analysis, which are listed below.

B. Consider a normal space l_1 of sequences $\alpha = \{\alpha\ (0),\ \alpha\ (1),\ \ldots\}$ of numbers with norm $|\ \alpha\ | = \sum_{i\geqslant 0} |\ \alpha\ (i)\ |$. We will write $\alpha \geqslant 0$, if $\alpha\ (i) \geqslant 0\ \forall\ i$. Similarly, for α and β from l_1, the inequalities $\alpha \geqslant \beta$ or $\beta \leqslant \alpha$ mean that $\alpha - \beta \geqslant 0$. If $\alpha \geqslant \beta$, and $\alpha \neq \beta$, then we will write $\alpha > \beta$ or $\beta < \alpha$.

We will employ the sign $\xrightarrow{w}$ to denote **weak** (coordinate-wise) convergence. E.g., $\alpha_n \xrightarrow{w} \alpha$ means that the sequence $\{\alpha_n\} \subset l_1$ converges to $\alpha \in l_1$ weakly (i.e., $\alpha_n\ (i) \to \alpha\ (i)\ \forall i$).

C. We call a linear operator $P : l_1 \to l_1$ **stochastic** if

(1) $\alpha \geqslant 0$ implies $P\alpha \geqslant 0$, while $|\ P\alpha\ | = |\ \alpha\ |$, and

(2) $|\ P\alpha\ | \leqslant |\ \alpha\ |$ for all $\alpha \in l_1$.

Consider a family $\mathcal{P} = \{P^t,\ t \in T\}$ of stochastic operators with the semigroup property $P^{s+t} = P^s \cdot P^t$ for all s and t from T, where $T = \{1,\ 2,\ \ldots\}$ or $T = (0,\ \infty)$.

A semigroup $\mathcal{P} = \{P^t,\ t \in T\}$ of stochastic operators is said to be **contracting** if it follows from $\alpha \in l_1$ and $|\ P^t\alpha\ | = |\ \alpha\ |$ for all $t \in T$ that $\alpha \geqslant 0$ or $\alpha \leqslant 0$.

A point $\alpha \in l_1$ is said to be **invariant** (with respect to $\mathcal{P}$) if $P^t\alpha = \alpha$ for all $t \in T$. The point $\alpha = 0$ is called the **trivial** invariant point.

D. Proposition 1. Let a semigroup $\mathcal{P} = \{P^t,\ t \in T\}$ of stochastic operators be contracting. Then:

1°. If there exists no non-trivial invariant point, $P^t\alpha \xrightarrow{w} 0$ as $t \to \infty$ for any $\alpha \in l_1$.

2°. If there exists a non-trivial invariant point, then there is a unique invariant point π satisfying the conditions $\pi \geqslant 0$ and $|\pi| = 1$. Meanwhile, for any $\alpha \in l_1$, $P^t\alpha \xrightarrow{w} \lambda_\alpha \cdot \pi$, where λ_α is a real number. If, however, $\alpha > 0$, then $0 < \lambda_\alpha \leqslant |\alpha|$.

Besides, an invariant point is unique up to a real factor.

E. We shall use only the following properties of the space l_1.

P1. $\alpha \geqslant 0$, $\beta \geqslant 0$ implies $|\alpha + \beta| = |\alpha| + |\beta|$.

P2. Any set bounded under the norm is weakly compact.

P3. $\alpha_n \xrightarrow{w} \alpha$, $|\alpha_n| \to c$, $\beta \in l_1$ implies

$$|\alpha_n - \alpha + \beta| \to c - |\alpha| + |\beta|.$$

P4. For any $\alpha \in l_1$, there exist unique elements $\alpha^+ \geqslant 0$ and $\alpha^- \geqslant 0$ such that $\alpha = \alpha^+ - \alpha^-$ and $|\alpha| = |\alpha^+| + |\alpha^-|$.

We shall only verify P3 of these properties, since the others are obvious.

For $\alpha \in l_1$ and $K \subset E = \{0, 1, 2, \ldots\}$, we put

$$|\alpha|_K = \sum_{i \in K} |\alpha(i)|.$$

Putting $K^* = E \setminus K$, we have

$$\begin{aligned} |\alpha_n - \alpha + \beta| &= |\alpha_n - \alpha + \beta|_K + |\alpha_n - \alpha + \beta|_{K^*} \\ &\leqslant |\alpha_n - \alpha|_K + |\beta|_K + |\alpha_n|_{K^*} + |\alpha|_{K^*} + |\beta|_{K^*} \\ &= |\alpha_n - \alpha|_K + |\beta| + |\alpha_n| - |\alpha_n|_K + |\alpha|_{K^*}. \end{aligned}$$

If K is a finite set, e.g., $K = \{0, 1, \ldots, N\}$, then

$$\overline{\lim}\, |\alpha_n - \alpha + \beta| \leqslant |\beta| + c - |\alpha|_K + |\alpha|_{K^*}.$$

Now, if $K \uparrow E$, e.g., $N \to +\infty$, then

$$\overline{\lim}\, |\alpha_n - \alpha + \beta| \leqslant c - |\alpha| + |\beta|.$$

Similarly,

$$\begin{aligned} |\alpha_n - \alpha + \beta| &\geqslant |\beta|_K - |\alpha_n - \alpha|_K + |\alpha_n|_{K^*} - |\alpha|_{K^*} - |\beta|_{K^*} \\ &= (|\beta|_K + |\alpha_n| - |\alpha|_K) - |\alpha_n - \alpha|_K - |\alpha|_{K^*} - |\beta|_{K^*}. \end{aligned}$$

We obtain by the same method that $\underline{\lim}\, |\alpha_n - \alpha + \beta| \geqslant c - |\alpha| + |\beta|$.

F. We call a point $\pi \in l_1$ **ergodic** if it is invariant (with respect to $\mathcal{P}$), $\pi \geqslant 0$, $|\pi| = 1$ and $P^t\alpha \xrightarrow{w} |\alpha| \cdot \pi$ for all $\alpha \geqslant 0$. By Property P3, the latter condition is equivalent to $|P^t\alpha - |\alpha| \cdot \pi| \to 0$ for all $\alpha \geqslant 0$. It is clear that if there exists an ergodic point, then it is unique. Note that if π is ergodic, then $P^t\alpha$ converges under the norm to $(|\alpha^+| - |\alpha^-|) \cdot \pi$ already for all $\alpha \in l_1$ as $t \to \infty$, where α^+ and α^- are selected relative to α according to P4.

The space m conjugate to l_1 is the space of bounded numerical sequences $w = \{w(0), w(1), \ldots\}$ with norm $\|w\| = \sup_{i \geqslant 0} |w(i)|$.

We call a functional $w \in m$ **positive** if

$$w(\alpha) = [w, \alpha] = \sum_{i \geqslant 0} w(i)\,\alpha(i) \geqslant 0$$

for all $\alpha \geqslant 0$, which is equivalent to $w(i) \geqslant 0\ \forall\, i$. For two elements w_1 and w_2 from m, the inequalities $w_1 \geqslant w_2$ or $w_2 \leqslant w_1$ mean that $w_1 - w_2$ is a positive functional, whereas $w_1 > w_2$ means that $w_1 \geqslant w_2$, and $w_1 \neq w_2$.

Further, let $(P^t)^* : m \to m$ be a bounded linear operator conjugate to $P^t : l_1 \to l_1$. Put $P^{*t} = (P^t)^*$. Then $\mathcal{P}^* = \{P^{*t},\ t \in T\}$ is a semigroup of operators (i.e., $P^{*s+t} = P^{*s} \cdot P^{*t}$ for all s and t from T).

We call a point $w \in m$ **invariant** (with respect to $\mathcal{P}^*$) if $P^{*t}w = w$ for all $t \in T$.

Proposition 2. Let there exist a non-trivial invariant point from l_1. Then the following statements are equivalent.

1°. There exists an ergodic point.

2°. (1) The semigroup $\mathcal{P}$ is contracting.

(2) An invariant point in m is unique up to a real factor.

G. A semigroup $\mathcal{P} = \{P^t,\ t \in T\}$ of stochastic operators is said to be **irreducible** if, for any $\alpha > 0$ and $w > 0$ from l_1 and m, respectively, it follows that there exists $t \in T$ such that $[w, P^t\alpha] > 0$.

Proposition 3. Let there exist a non-trivial invariant point from l_1. Then the following statements are equivalent.

1°. A semigroup $\mathcal{P}$ is contracting and irreducible.

2°. There exists an ergodic point π such that $[w, \pi] > 0$ for all $w > 0$ (i.e., $\pi = \{\pi_j\}$ and $\pi_j > 0\ \forall\, j$).

H. For a homogeneous Markov chain with transition probability matrix $\{p_{ij}^t\}$, $t \in T$, we define the semigroup $\mathcal{P} = \{P^t,\ t \in T\}$ of stochastic operators, acting in l_1, by the equalities

$$(P_\alpha^t)(j) = \sum_{i \geqslant 0} \alpha(i)\, p_{ij}^t, \quad j = 0, 1, 2, \ldots.$$

Then Theorems 1-3 correspond to Propositions 1-3. Meanwhile, we only have to verify the validity of the following statements.

(a) The contractibility of the semigroup $\mathcal{P}$ is equivalent to that of the Markov chain.

(b) The irreducibility of the semigroup $\mathcal{P}$ is equivalent to that of the Markov chain.

Let us prove statement (a). Let a semigroup $\mathcal{P}$ be contracting, and (i_1, i_2) a certain pair of states, with $i_1 \neq i_2$. Put $\varepsilon_i = (0, \ldots, 0, 1, 0)$, where unity is in the i-th place, and $i \geqslant 0$. Let $\alpha = \varepsilon_{i_1} - \varepsilon_{i_2} \in l_1$. We have $|\alpha| = 2$. If, however, the chain is not contracting, then

$$|P_\alpha^t| = \sum_j |p_{i_1 j}^t - p_{i_2 j}^t| = \sum_j (p_{i_1 j}^t + p_{i_2 j}^t) = 2$$

for all $t \in T$, giving $|P_\alpha^t| = |\alpha| \ \forall\, t$, i.e., $\alpha \geqslant 0$ or $\alpha \leqslant 0$, which is not true.

Now, conversely, let a chain be contracting, and α any element from l_1, so that, for certain i_1 and i_2, the inequalities $\alpha(i_1) > 0$ and $\alpha(i_2) < 0$ hold. It suffices to see that $|P^t \alpha| < |\alpha|$ for a certain $t \in T$. Indeed, there exist j_0 and t such that $p_{i_1 j_0}^t > 0$, and $p_{i_2 j_0}^t > 0$. But then

$$|P_\alpha^t| = \sum_j \Big|\sum_i \alpha(i)\, p_{ij}^t\Big| < \sum_i \sum_j |\alpha(i)\, p_{ij}^t| = \sum_i |\alpha(i)| = |\alpha|.$$

Statement (b) is obvious. It only should be taken into account that

$$[w, P_\alpha^t] = \sum_{i,j \geqslant 0} \alpha(i)\, w(j)\, p_{ij}^t.$$

Remark. Propositions 1-3 remain valid if l_1 and m are replaced by a normed space M and the space M^* conjugate to it, respectively. In addition, a convex cone M^+ determining an ordering in M by the condition "$\alpha \geqslant 0$ means that $\alpha \in M^+$," and a certain convergence called weak (along with strong convergence generated by the norm of the space M) are given.

The choice of the cone M^+, norm in M, weak convergence and semigroup $\mathcal{P}$ is determined by P1-P4.

§ 5. Law of Large Numbers for Markov Chains

We shall use the previous notation and definitions, applying them to a Markov chain $\{\xi(t) : t \in T\}$. In the case $T = (0, \infty)$, the homogeneous Markov chain is said to be **standard** if $p_{ij}^t \to \delta_{ij}$ when $t \downarrow 0$ for all i and j from E.

Let $k \in E$, and $x_k(t)$ be the indicator of the event $\{\xi(t) = k\}$. Put

$$X_k(t) = \frac{1}{t}[x_k(1) + \ldots + x_k(t)] \quad \text{if } T = \{1, 2, \ldots\},$$

and

$$X_k(t) = \frac{1}{t}\int_0^t x_k(\tau)\, d\tau \quad \text{if } T = (0, \infty).$$

Lebesgue integrability is thereby assumed on any finite interval of (almost all) trajectories of the process $\{x_k(t) : t \in T\}$.

The goal of the section is to prove the following statement.

Proposition (LLN for Markov chains). Let, for a homogeneous Markov chain and each $k \in E$, there exist $\lim_{t\to\infty} p_{ik}^t = \pi_k$ independent of $i \in E$. Besides, in the case $T = (0, \infty)$, we will assume that the chain is standard. Then, for any $\lambda > 0$,

$$E\,|X_k(t) - \pi_k|^\lambda \to 0, \quad k \in E,$$

as $t \to +\infty$.

Proof. We assume that the conditions of the proposition are fulfilled, and put $P_{ij}(t) = p_{ij}^t$.

1°. It suffices to prove that

$$E\,|X_k(t) - \pi_k|^2 \to 0 \text{ as } t \to \infty.$$

Indeed, putting

$$\eta = \eta(t) = X_k(t) - \pi_k,$$

we obtain $|\eta| \leqslant 1$, and

$$E\,|\eta|^\lambda \leqslant \begin{cases} E\,|\eta|^2 & \text{when } \lambda \geqslant 2 \\ E^{\lambda/2}\,|\eta|^2 & \text{when } 0 < \lambda \leqslant 2 \end{cases}.$$

When $0 < \lambda \leqslant 2$, this relation coincides with the Liapunov inequality

$$E^{1/\lambda}\,|\eta|^\lambda \leqslant E^{1/2}\,|\eta|^2.$$

2°. Let $T = \{1, 2, \ldots\}$. Then

$$\operatorname{var} X_k(t) = E\,[X_k(t) - EX_k(t)]^2$$

$$= \frac{1}{t^2} E \sum_{u=1}^{t} \sum_{v=1}^{t} [x_k(u) - P_{ik}(u)]\,[x_k(v) - P_{ik}(v)]$$

$$= \frac{1}{t^2} \sum_{u=1}^{t} \sum_{v=1}^{t} E\,[x_k(u) - P_{ik}(u)]\, x_k(v)$$

$$= \frac{2}{t^2} \sum_{1 \leqslant u < v \leqslant t} E\,[x_k(u) - P_{ik}(u)]\, x_k(v) + \frac{1}{t^2} \sum_{u=1}^{t} P_{ik}(u)\,[1 - P_{ik}(u)]$$

$$= \frac{2}{t^2} \sum_{1 \leqslant u < v \leqslant t} P_{ik}(u)\,[P_{kk}(v-u) - P_{ik}(v)] + \frac{1}{t^2} \sum_{u=1}^{t} P_{ik}(u)\,[1 - P_{ik}(u)]$$

$$= \frac{2}{t^2} \sum_{u=1}^{t} P_{ik}(u) \sum_{w=1}^{t-u} [P_{kk}(w) - P_{ik}(u+w)] + \frac{1}{t^2} \sum_{u=1}^{t} P_{ik}(u)\,[1 - P_{ik}(u)].$$

Since

$$\lim_{w\to\infty} P_{hk}(w) = \lim_{w\to\infty} P_{ik}(w) = \pi_k,$$

it follows that

$$\operatorname{var} X_k(t) \to 0 \text{ as } t\to\infty.$$

3°. It remains to notice for the case $T=\{1, 2, \ldots\}$ that

$$E\,|\,X_k(t)-\pi_k\,|^2 = \operatorname{var} X_k(t) + [EX_k(t)-\pi_k]^2,$$

and, as $t\to\infty$,

$$EX_k(t) = \frac{1}{t}\sum_{u=1}^{t} Ex_k(u) = \frac{1}{t}\sum_{u=1}^{t} P_{ik}(u) \to \pi_k.$$

4°. Now, let $T=(0, \infty)$. Then, as $t\to\infty$,

$$\begin{aligned}
\operatorname{var} X_k(t) &= \frac{1}{t^2} E \int_0^t \int_0^t [x_k(u)-P_{ik}(u)]\,[x_k(v)-P_{ik}(v)]\,du\,dv \\
&= \frac{2}{t^2} \iint\limits_{0\leqslant u\leqslant v\leqslant t} E\,[x_k(u)-P_{ik}(u)]\,[x_k(v)-P_{ik}(v)]\,du\,dv \\
&= \frac{2}{t^2} \iint\limits_{0\leqslant u\leqslant v\leqslant t} E\,[x_k(u)-P_{ik}(u)]\,x_k(v)\,du\,dv \\
&= \frac{2}{t^2} \iint\limits_{0\leqslant u\leqslant v\leqslant t} P_{ik}(u)\,[P_{kk}(v-u)-P_{ik}(v)]\,du\,dv \\
&= \frac{2}{t^2} \int_0^t P_{ik}(u)\,du \int_0^{t-u} [P_{kk}(w)-P_{ik}(u+w)]\,dw \to 0.
\end{aligned}$$

5°. Again, it remains to notice that

$$E\,|\,X_k(t)-\pi_k\,|^2 = \operatorname{var} X_k(t) + [EX_k(t)-\pi_k]^2,$$

and, as $t\to\infty$,

$$EX_k(t) = \frac{1}{t}\int_0^t Ex_k(u)\,du = \frac{1}{t}\int_0^t P_{ik}(u)\,du \to \pi_k.$$

6°. Note that, for any i and j from E, the function $P_{ij}(t)$ is continuous (and thereby integrable). In fact, it suffices to verify that

$$|\,P_{ij}(t+h)-P_{ij}(t)\,| \leqslant 2[1-P_{ii}(|\,h\,|)].$$

For $h>0$, we have

$$\begin{aligned}
|\,P_{ij}(t+h)-P_{ij}(t)\,| &= \Big|\sum_k P_{ik}(h)\,P_{kj}(t)-P_{ij}(t)\Big| \\
&= \Big|[P_{ii}(h)-1]\,P_{ij}(t)+\sum_{k\neq i} P_{ik}(h)\,P_{kj}(t)\Big| \\
&\leqslant |\,P_{ii}(h)-1\,|\,P_{ij}(t)+1-P_{ii}(h) \leqslant 2\,[1-P_{ii}(h)].
\end{aligned}$$

Note that, actually,

$$|\,P_{ij}(t+h)-P_{ij}(t)\,| \leqslant [1-P_{ii}(|\,h\,|)].$$

By the way, if we do not require that the process should be standard (continuous at zero), then the functions $P_{ij}(t)$ may even be non-measurable.

§ 6. Structure of Homogeneous Continuous-Time Markov Process with Finite State Space

In this section, we shall assume that $T=(0, \infty)$, and $|\,E\,|<\infty$, i.e., consider only homogeneous continuous-time Markov processes with finite state space E. For brevity, such processes will be simply called Markov processes. Besides, we shall use the notation and definitions of the previous sections.

We call two Markov processes **equivalent** if the transition probabilities corresponding to them coincide.

As it turns out, an arbitrary standard Markov process (continuous at zero) is equivalent to a minimal process (see Example 4, § 2). The goal of this section is to formulate the corresponding statements which will at the same time indicate a method of determining probabilities $P_{ij}(t)$ in terms of simpler parameters of the process (in terms of α_i and q_{ij}).

Proposition 1. For a standard Markov process, there exists

$$\lim_{t\downarrow 0} \frac{1}{t}(P^t-I) = A = \{a_{ij}\},$$

where I is the unit matrix, and the elements of the matrix A are finite numbers satisfying the condition

$$a_{ij}\geqslant 0 \text{ when } i\neq j, \quad \sum_j a_{ij} = 0. \tag{1}$$

The operator A is called the **infinitesimal operator** of a Markov process. Note that the indicated limit is understood as a limit applied to each element of the matrix.

Proposition 2. The elements of the matrix $P^t = P(t) = \{P_{ij}(t)\}$ are differentiable at any point $t\geqslant 0$, and

$$P'(t) = A\cdot P(t) = P(t)\cdot A, \quad P(0) = I \tag{2}$$

Proposition 3. If a matrix A satisfies condition (1), then equation (2) has a unique solution $P(t) = \{P_{ij}(t)\}$, $t\geqslant 0$, and this solution satisfies the conditions (1) $P(s+t) = P(s)\cdot P(t)$, (2) $P_{ij}(t)\geqslant 0$, (3) $\sum_j P_{ij}(t) = 1$, (4) $P_{ij}(t)\to\delta_{ij}$ when $t\downarrow 0$ and (5) $\lim_{t\downarrow 0}\frac{1}{t}[P(t)-I] = A$.

Proposition 4. A minimal Markov process is standard. The infinitesimal matrix $A = \{a_{ij}\}$ of a minimal Markov process without absorbing states is given by the relations

$$a_{ij} = \alpha_i q_{ij} \text{ when } i \neq j, \quad a_{ii} = -\alpha_i.$$

Proposition 5. A state $i \in E$ is said to be **absorbing** if $P_{ij}(t) = \delta_{ij}$ for all j and t. A state i is absorbing if and only if $a_{ii} = 0$. A standard Markov process with infinitesimal operator $A = \{a_{ij}\}$ and without absorbing states is equivalent to a minimal Markov process for which

$$\alpha_i = -a_{ii}, \quad q_{ij} = \frac{a_{ij}}{\alpha_i} \text{ when } i \neq j, \quad q_{ii} = 0.$$

Proof of Proposition 1. A. Let a numerical positive semi-definite function f defined and continuous on $(0, \infty)$ satisfy the conditions

$$\forall \ \varepsilon > 0 \ \exists \ t_0 > 0 \text{ such that } f(nh) \geqslant (1 - \varepsilon) nf(h)$$

for $h > 0$, $nh \leqslant t_0$, where n is a positive number. In this case, we show that

$$\lim_{h \downarrow 0} \frac{f(h)}{h}$$

exists, and is finite.

In fact, let

$$\overline{\lim_{h \downarrow 0}} \frac{f(h)}{h} = c.$$

It follows from the inequality

$$\frac{f(nh)}{nh} \geqslant (1 - \varepsilon) \frac{f(h)}{h}$$

and continuity of f on $(0, \infty)$ that $c < +\infty$. Therefore, there exists a number t_1 such that $0 < t_1 < \frac{t_0}{2}$ and

$$\frac{f(t_1)}{t_1} \leqslant c + \varepsilon.$$

But then there exists $h_0 > 0$ such that

$$\frac{f(t)}{t} \leqslant c + 2\varepsilon \text{ when } |t - t_1| < h_0.$$

Now, let $0 < h < \min(h_0, t_0)$. Take such an integer $n \geqslant 1$ that $t_1 \leqslant nh < t_1 + h \leqslant t_0$. Then

$$(1 - \varepsilon) \frac{f(h)}{h} \leqslant \frac{f(nh)}{nh} \leqslant c + 2\varepsilon.$$

Due to the arbitrariness of $\varepsilon > 0$, this gives

$$\overline{\lim_{h \downarrow 0}} \frac{f(h)}{h} \leqslant c.$$

B. We now show that if $i \neq j$, the function $f(t) = P_{ij}(t)$ satisfies the condition of Item A.

For any $h > 0$, consider a homogeneous discrete-time Markov chain with one-step transition probability matrix $P(h)$. Let

$f^n_{ij}(h)$ be the probability of transition from i into j precisely in n steps,

and

$f^n_{i/j(n)}(h)$ the probability of transition from i into i in n steps without transitions into a state j.

Then

$$P_{ij}(nh) \geqslant \sum_{k=0}^{n-1} f^h_{i/j}(h) P_{ij}(h) P_{jj}((n-k-1)h), \tag{3}$$

$$P_{ii}(kh) = f^k_{i/j}(h) + \sum_{m=1}^{k-1} f^m_{ij}(h) P_{ji}((k-m)h), \tag{4}$$

and

$$\sum_{m=1}^{k-1} f^m_{ij}(h) \leqslant 1. \tag{5}$$

Taking into account the continuity of the functions $P_{ij}(t)$ (because the Markov process is standard; see the end of § 5), we derive that, for any $\varepsilon > 0$, there exists $t_0 > 0$ such that

$$P_{ij}(nh) \geqslant (1 - \varepsilon) nP_{ij}(h)$$

when $h > 0$, $nh \leqslant t_0$, and n is a positive integer.

In fact, we obtain from (4)-(5) that

$$f^k_{i/j}(h) \geqslant P_{ii}(kh) - \max_{1 \leqslant m < k} P_{ji}((k-m)h).$$

Further, for $\delta > 0$, there exists $t_0 > 0$ such that if $0 \leqslant t < t_0$, the following inequalities hold, viz.,

$$P_{ji}(t) \leqslant \delta, \quad P_{ii}(t) \geqslant 1 - \delta, \quad P_{jj}(t) \geqslant 1 - \delta,$$

whence $f^k_{i/j}(h) \geqslant 1 - 2\delta$, and when $nh \leqslant t_0$, we obtain from (3) that

$$P_{ij}(nh) \geqslant (1 - 2\delta) \sum_{k=0}^{n-1} P_{ij}(h)(1 - \delta) \geqslant (1 - 3\delta) nP_{ij}(h).$$

It remains to put $\delta = \varepsilon/3$.

C. It follows from A and B that if $i \neq j$,

$$\lim_{t \downarrow 0} \frac{P_{ij}(t)}{t} = a_{ij} \geqslant 0$$

exists and is finite. And since

$$\frac{1}{t}[1 - P_{ii}(t)] = \sum_{j \neq i} \frac{P_{ij}(t)}{t}$$

when $t > 0$, and the number of addends is finite in the latter sum,

$$\lim_{t \downarrow 0} \frac{1}{t} [P_{ii}(t) - 1] = a_{ii}$$

exists and is finite, with $\sum_j a_{ij} = 0$.

Proof of Proposition 2. When $t = 0$, the statement coincides with Proposition 1 and $P(0) = I$. Now, let $t > 0$, and $h > 0$. Then

$$P^{t+h} - P^t = P^t (P^h - I), \text{ and } P^{t-h} - P^t = -P^{t-h}(P^h - I).$$

Since there exists

$$\lim_{h \downarrow 0} \frac{1}{h} (P^h - I) = A,$$

and P^t is continuous, P^t is differentiable. It follows from the equalities

$$\frac{1}{h}(P^{t+h} - P^t) = P^t \cdot \frac{1}{h}(P^h - I) = \frac{1}{h}(P^h - I) \cdot P^t$$

that

$$P'(t) = P(t) \cdot A = A \cdot P(t).$$

Proof of Proposition 3. Note that a unique solution of equation (2) is of the form

$$P(t) = e^{tA}.$$

We shall only verify properties (2) and (3), since the other are obvious.

A. Put

$$\lambda = \min_{i \in E} a_{ii} \leqslant 0,$$

$$Q(t) = e^{-\lambda t} P(t).$$

Then

$$Q'(t) = -\lambda e^{-\lambda t} P(t) + e^{-\lambda t} P'(t)$$
$$= -\lambda Q(t) + e^{-\lambda t} P(t) \cdot A = -\lambda Q(t) + Q(t) \cdot A,$$

or

$$Q'(t) = Q(t) \cdot B, \quad Q(0) = I, \quad B = A - \lambda I.$$

Since

$$Q(t) = e^{tB} = \sum_{n \geqslant 0} \frac{t^n}{n!} B^n \text{ and } b_{ij} = a_{ij} - \lambda \delta_{ij} \geqslant 0,$$

$Q_{ij}(t) \geqslant 0$; therefore, $P_{ij}(t) \geqslant 0$.

B. We derive from

$$P'_{ij}(t) = \sum_k P_{ik}(t)\, a_{kj}$$

that

$$[\sum_j P_{ij}(t)]' = \sum_j P'_{ij}(t) = \sum_k P_{ik}(t) \cdot \sum_j a_{kj} = 0,$$

whence

$$\sum_j P_{ij}(t) = \text{const} = \sum_j P_{ij}(0) = 1.$$

Note that if we had $\sum_j a_{kj} = 0 \ \forall k$ instead of $\sum_j a_{kj} \leqslant 0 \ \forall k$ (and, certainly, $a_{ij} \geqslant 0$ when $i \neq j$), then we should have

$$[\sum_j P_{ij}(t)]' \leqslant 0;$$

therefore,

$$\sum_j P_{ij}(t) \leqslant \sum_j P_{ij}(0) = 1.$$

The simple proofs of the remaining propositions are left to the reader.

Example. Consider a birth-and-death process $\{\xi(t) : t \geqslant 0\}$ with finite set space $E = \{0, 1, \ldots, N\}$. Recall that this implies the sojourn at a state $i \in E$ having the exponential distribution with parameter α_i (and not depending on the previous trajectory of the process) and subsequently, the process changing over to the neighbouring state $i + 1$ (if $i < N$) or $i - 1$ (if $i > 0$) with probabilities p_i and $q_i = 1 - p_i$, respectively, whereas $p_0 = 1$, $q_N = 1$. Assume that $0 < p_i < 1$ when $0 < i < N$.

Put

$$P_k(t) = P\{\xi(t) = k\}, \quad t \geqslant 0, \quad k \in E,$$

and show that

(1)
$$P'_0(t) = -a_0 P_0(t) + b_1 P_1(t),$$
$$P'_k(t) = a_{k-1} P_{k-1}(t) - (a_k + b_k) P_k(t) + b_{k+1} P_{k+1}(t), \quad 0 < k < N,$$
$$P'_N(t) = a_{N-1} P_{N-1}(t) - b_N P_N(t),$$

where $a_i = \alpha_i p_i$, $b_i = \alpha_i q_i$,

(2) there exist limits

$$\lim_{t \to \infty} P_k(t) = \pi_k, \quad k \in E,$$

not depending on the initial state $\xi(0) \in E$, and

(3)
$$\pi_k = \rho_k \pi_0, \quad \sum_{k=0}^{N} \pi_k = 1, \text{ where } \rho_0 = 1,$$
$$\rho_k = \frac{a_0 \ldots a_{k-1}}{b_1 \ldots b_k}, \quad 1 \leqslant k \leqslant N.$$

In fact, statement (1) follows from Propositions 2 and 4, while (2) is a consequence of Theorem 0, § 3, and the chain under consideration being contracting.

We now verify statement (3). It follows from statement (2) and system of differential equations (1) that there exist finite limits $\lim_{t\to\infty} P_k'(t) = c_k$. If we had $c_k \neq 0$, then the function $P_k(t)$ would be unbounded. Therefore, $c_k = 0\ \forall\, k$, and

$$0 = -a_0\pi_0 + b_1\pi_1,$$
$$0 = a_{k-1}\pi_{k-1} - (a_k + b_k)\pi_k + b_{k+1}\pi_{k+1}, \quad 0 < k < N,$$
$$0 = a_{N-1}\pi_{N-1} - b_N\pi_N.$$

Putting

$$z_k = b_k\pi_k - a_{k-1}\pi_{k-1}, \quad 1 \leqslant k \leqslant N,$$

we obtain

$$z_1 = 0, \quad z_{k+1} - z_k = 0 \quad \text{when} \quad 0 < k < N, \quad z_N = 0,$$

i.e., $z_k = 0\ \forall k$, whence

$$\pi_k = \frac{a_{k-1}}{b_k}\pi_{k-1}.$$

It remains to make use of the condition $\sum \pi_k = 1$.

§ 7. Structure of Homogeneous Continuous-Time Markov Process with Countable State Space

We retain the same notation and definitions. In this case, the state space is $E = \{0, 1, 2, \ldots\}$.

Proposition 1. For a standard Markov process, when $i \neq j$,

$$\lim_{t \downarrow 0} \frac{P_{ij}(t)}{t} = a_{ij} \geqslant 0 \tag{1}$$

exists and is finite, whereas

$$\lim_{t \downarrow 0} \frac{1}{t}[1 - P_{ii}(t)] = \alpha_i = -a_{ii} \tag{2}$$

exists, but may be infinite. The elements of the matrix $A = \{a_{ij}\}$ satisfy the condition

$$a_{ij} \geqslant 0 \text{ when } i \neq j, \qquad \sum_j a_{ij} \leqslant 0\ \forall i. \tag{3}$$

The operator (matrix) A is called the **infinitesimal operator** of the Markov process.

If the matrix A satisfies the condition

$$a_{ij} \geqslant 0 \text{ when } i \neq j, \quad \sum_j a_{ij} = 0 \text{ and } \alpha_i = -a_{ii} < \infty\ \forall i,$$

then the Markov process is said to be **conservative**, in which case the matrix A is also called conservative.

According as which of the relations

$$\alpha_i = \infty, \quad \alpha_i = 0 \text{ or } \quad 0 \leqslant \alpha_i < \infty$$

is fulfilled, the state i is said to be **instantaneous**, **absorbing** or **stable**, respectively. In this section, all states are assumed to be stable.

Proposition 2. For a conservative Markov process, the transition probabilities $P_{ij}(t)$ are differentiable, and satisfy the Kolmogorov backward equation

$$P'(t) = A \cdot P(t), \quad P(0) = I, \tag{4}$$

where $I = \{\delta_{ij}\}$, $P(t) = \{P_{ij}(t)\}$.

We call a matrix function $t \to Q(t) = \{Q_{ij}(t)\}$ **stochastic** if, for any t, the matrix $Q(t)$ is stochastic, i.e.,

$$Q_{ij}(t) \geqslant 0 \text{ and } \sum_j Q_{ij}(t) = 1.$$

If, instead of the latter equality, the following inequality holds, viz.,

$$\sum_j Q_{ij}(t) \leqslant 1,$$

then a **semi-stochastic matrix** and **function** are spoken of.

Proposition 3. If a matrix $A = \{a_{ij}\}$ of finite numbers satisfies condition (3), then, in the class of semi-stochastic matrix functions, there exists a function $P(t)$ which

(1) is a solution of the Kolmogorov backward equation (4),

(2) is a solution of the Kolmogorov forward equation

$$P'(t) = P(t) \cdot A, \quad P(0) = I, \tag{5}$$

and

(3) satisfies the Chapman-Kolmogorov equation

$$P(s + t) = P(s) \cdot P(t).$$

Proposition 4. Let a matrix $A = \{a_{ij}\}$ of finite numbers satisfy condition (3). Then the Kolmogorov backward equation (4) has a unique solution in the class of semi-stochastic matrix functions, which is a stochastic matrix function if and only if the same is true for Kolmogorov forward equation (5).

Proof of Proposition 1. The proof of the existence and finiteness of limit (1) is the same as in the case of a finite state space (see the proof of Proposition 1 in the previous section).

We now prove the existence of limit (2). First, we show that it follows from the Markov process being standard that

$$P_{ii}(t) > 0 \ \forall\, i \in E,\ t \geqslant 0$$

(at the point $t = 0$, the function $P_{ii}(t)$ is defined by continuity by the value 1). In fact, it follows from $P_{ii}(t) \to 1$ that, when $t \downarrow 0$, there exists a number $h > 0$ such that $P_{ii}(t) > 0$ when $0 \leqslant t \leqslant h$. For an arbitrary $t \geqslant 0$ and a positive integer n such that $\frac{t}{n} \leqslant h$,

$$P_{ii}(t) \geqslant \left[P_{ii}\left(\frac{t}{n}\right)\right]^n > 0.$$

Put $g(t) = -\ln P_{ii}(t)$. It follows from the inequality

$$P_{ii}(s+t) \geqslant P_{ii}(s) \cdot P_{ii}(t) \text{ for } s \geqslant 0,\ t \geqslant 0$$

that

$$g(x+y) \leqslant g(x) + g(y) \text{ for } x \geqslant 0, \quad y \geqslant 0.$$

Besides, $g(x) \geqslant 0$ when $x \geqslant 0$, and $g(x) \to 0$ when $x \downarrow 0$. If we show that there exists

$$\lim_{x \downarrow 0} \frac{g(x)}{x}, \tag{6}$$

then it will follow from the equalities

$$\lim_{t \downarrow 0} \frac{g(t)}{t} = \lim_{t \downarrow 0} \frac{-\ln\{1-[1-P_{ii}(t)]\}}{t} = \lim_{t \downarrow 0} \frac{1-P_{ii}(t)}{t}$$

that limit (2) exists, due to which we put

$$\sup_{x>0} \frac{g(x)}{x} = \lambda.$$

If $\lambda < \infty$, then, for $\varepsilon > 0$, there exists $x_0 > 0$ such that

$$\lambda - \varepsilon \leqslant \frac{g(x_0)}{x_0}.$$

Let $0 < x \leqslant x_0$, and $x_0 = nx + h$, where $0 \leqslant h < x$, and n is a positive integer. Then

$$\lambda - \varepsilon \leqslant \frac{g(x_0)}{x_0} \leqslant \frac{ng(x)+g(h)}{x_0} = \frac{nx}{x_0} \cdot \frac{g(x)}{x} + \frac{g(h)}{x_0} \leqslant \frac{g(x)}{x} + \frac{g(h)}{x_0},$$

whence, due to $g(h) \to 0$ as $x \to 0$, we have

$$\lambda - \varepsilon \leqslant \lim_{x \downarrow 0} \frac{g(x)}{x},$$

which, along with the inequality

$$\overline{\lim_{x \downarrow 0}} \frac{g(x)}{x} \leqslant \lambda$$

implies the existence of limit (6).

It remains to verify (3). This inequality follows from

$$\sum_{j \neq i} \frac{P_{ij}(h)}{h} + \frac{P_{ii}(h)-1}{h} = 0 \quad \forall\, i,$$

the existence of the limit of each addend when $h \downarrow 0$, and Fatou lemma. More simply, (3) follows from the inequality

$$\sum_{i \neq j \in M} \frac{P_{ij}(h)}{h} + \frac{P_{ii}(h)-1}{h} \leqslant \frac{1}{h}\left[\sum_j P_{ij}(h) - 1\right] = 0$$

for $a_i < \infty$, where M is a finite state space. In fact, when $h \downarrow 0$, we derive that

$$\sum_{j \in M} a_{ij} \leqslant 0,$$

and, due to the arbitrariness of the choice of M and $a_{ij} \geqslant 0$ when $i \neq j$, we obtain (3).

Proof of Proposition 2. Let i and j be two fixed states, $h > 0$, and M an arbitrary finite state space of i, $j \in M$. By the Chapman-Kolmogorov equation,

$$\frac{1}{h}[P_{ij}(t+h) - P_{ij}(t)] = \sum_{k \in M} \frac{1}{h}[P_{ik}(h) - \delta_{ik}]\, P_{kj}(t) + \Delta(h, M), \tag{7}$$

$$0 \leqslant \Delta(h, M) = \frac{1}{h} \sum_{k \in M} P_{ik}(h) \cdot P_{kj}(t) \leqslant \frac{1}{h} \sum_{k \in M} P_{ik}(h)$$

$$= \frac{1}{h}\left[1 - \sum_{k \in M} P_{ik}(h)\right].$$

For $\varepsilon > 0$, we select M so that

$$\sum_{k \in M} a_{ik} \leqslant \varepsilon,$$

and, having chosen M, we take a number $h_0 > 0$ so that

$$\left|\frac{1}{h} P_{ik}(h) - a_{ik}\right| \leqslant \frac{\varepsilon}{|M|} \text{ when } k \neq i, \text{ and } k \in M,$$

$$\left|\frac{1}{h}[1 - P_{ii}(h)] - \alpha_i\right| \leqslant \frac{\varepsilon}{|M|}$$

for $0 < h \leqslant h_0$. Hence,

$$0 \leqslant \Delta(h, M) \leqslant 2\varepsilon \quad \text{when } 0 < h \leqslant h_0.$$

If, however, $h < 0$, then this case is reduced to the previous if we assume $h > 0$, and replace t by $t - h$. Meanwhile, we have to take into account in (7) that $P_{kj}(t-h) \to P_{kj}(t)$ when $h \downarrow 0$.

Thus,

$$\sum_{k\in M} a_{ik}P_{kj}(t) \leqslant \underline{\lim_{h\to 0}} \frac{1}{h}[P_{ij}(t+h)-P_{ij}(t)]$$

$$\leqslant \overline{\lim_{h\to 0}} \frac{1}{h}[P_{ij}(t+h)-P_{ij}(t)] \leqslant \sum_{k\in M} a_{ik}P_{kj}(t)+2\varepsilon,$$

whence

$$P'_{ij}(t)=\sum_k a_{ik}P_{kj}(t) \quad \forall\, i,\ j,$$

which coincides with (4) in matrix form.

Proof of Proposition 3. The idea of proof consists in considering a sequence of "truncated" Kolmogorov backward equations and studying the limiting behaviour of solutions to "truncated" equations.

More precisely, for each integer $n \geqslant 0$,we consider the "truncated Kolmogorov backward equation"

$$P'(t,\ n) = A(n)\cdot P(t,\ n),\ P(0,\ n) = I_n, \tag{8}$$

where

$$A(n)=\{a_{ij}(n)\},\ P(t,\ n)=\{P_{ij}(t,\ n)\},$$
$$I_n=\{\delta_{ij}\},\ i,\ j=0,\ 1,\ \ldots,\ n.$$

Note that the elements of the matrix $A(n)$ satisfy the condition

$$a_{ij}(n) \geqslant 0 \text{ when } i \neq j \text{ and } \sum_{j=0}^{n} a_{ij}(n) \leqslant 0\ \forall i. \tag{9}$$

It is obvious that a solution of equation (8) is unique, of the form

$$P(t,\ n) = e^{t\cdot A(n)},$$

and satisfies the truncated Kolmogorov forward equation

$$P'(t,\ n) = P(t,\ n)\cdot A(n), \quad P(0,\ n) = I_n \tag{10}$$

as well as the Chapman-Kolmogorov equation

$$P(s+t,\ n) = P(s,\ n)\cdot P(t,\ n). \tag{11}$$

We now show consecutively that

(A) $P(t,\ n)$ is a semi-stochastic matrix,

(B) $\forall\ i,\ j,\ t$, the sequence $P_{ij}(t,\ n)$ is not decreasing with respect to n; therefore, there exists $\lim\limits_{n\to\infty} P_{ij}(t,\ n) = Q_{ij}(t)$, and

(C) the matrix function $t \to Q(t) = \{Q_{ij}(t)\}$ is

1. semi-stochastic,
2. a solution of the Kolmogorov forward and backward equations

$$Q'(t) = Q(t)\cdot A, \quad Q'(t) = A\cdot Q(t), \quad Q(0) = I,$$

and

3. satisfies the Chapman-Kolmogorov equation

$$Q(s+t) = Q(s)\cdot Q(t).$$

A. The proof of the fact that the matrix $P(t,\ n)$ is semi-stochastic fully coincides, taking into account (9), with that for the case of a finite number of states (see the proof of Proposition 3, § 6).

B. Let $n \geqslant i$, $n \geqslant j$. After replacing n by $n+1$, we obtain from equation (10)

$$P'_{ij}(t,\ n+1) = \sum_{k=0}^{n+1} P_{ik}(t,\ n+1)\, a_{kj}$$

$$= \sum_{k=0}^{n} P_{ik}(t,\ n+1)\, a_{kj}(n) + P_{in+1}(t,\ n+1)\, a_{n+1j},$$

or, in matrix form,

$$W'(t) = W(t)\cdot A(n) + B(t),\ W(0) = I_n,$$

where

$$W(t) = \{W_{ij}(t) = P_{ij}(t,\ n+1):\ i,\ j = 0,\ 1,\ \ldots,\ n\},$$
$$B(t) = \{B_{ij}(t) = P_{in+1}(t,\ n+1)\cdot a_{n+1j} : i,\ j = 0,\ 1,\ \ldots,\ n\},$$

whence

$$W(t) = P(t,\ n) + \int_0^t P(t-u,\ n)\cdot B(u)\, du,$$

and, since $B_{ij}(t) \geqslant 0$,

$$W_{ij}(t) \geqslant P_{ij}(t,\ n), \quad \text{i.e.,} \quad P_{ij}(t,\ n+1) \geqslant P_{ij}(t,\ n).$$

C. We shall make use of the following simple statement.

Lemma. Let, for each number $i = 0,\ 1,\ 2,\ \ldots$, the sequence $\{c_i(n) : n = 0,\ 1,\ 2,\ \ldots\}$ be non-decreasing, and $\lim\limits_{n\to\infty} c_i(n) = c_i < \infty$. Then

$$\sum_{i\geqslant 0} c_i = \lim_{n\to\infty} \sum_{i\geqslant 0} c_i(n).$$

1. Since

$$P_{ij}(t,\ n) \geqslant 0,\ \sum_j P_{ij}(t,\ n) \leqslant 1, \quad Q_{ij}(t) = \lim_{n\to\infty} P_{ij}(t,\ n),$$

we infer that $Q_{ij}(t) \geqslant 0$, and, putting $c_j(n) = P_{ij}(t,\ n)$, obtain from the lemma that $\sum\limits_j Q_{ij}(t) \leqslant 1$, i.e., the matrix $Q(t)$ is semi-stochastic. Note that we have assumed $P_{ij}(t,\ n) = 0$ for $i > n$, and/or $j > n$.

2. First, we note that

$$Q_{ij}(0) = \lim_{n\to\infty} P_{ij}(0,\ n) = \delta_{ij}, \text{ i.e., } Q(0) = I.$$

It suffices to verify that $Q(t)$ satisfies the Kolmogorov backward and forward equations which are written in integral form

$$Q_{ij}(t) = \delta_{ij} + \int_0^t \Big(\sum_k a_{ik} Q_{kj}(u) \Big)\, du,$$

(12)

$$Q_{ij}(t) = \delta_{ij} + \int_0^t \Big(\sum_k Q_{ik}(u) \cdot a_{kj} \Big)\, du,$$

or to verify the validity of the first of these and the equality

$$\sum_k a_{ik} Q_{kj}(t) = \sum_k Q_{ik}(t) \cdot a_{kj} \quad \forall\, i,\ j,\ t. \tag{13}$$

Put

$$\alpha_{ij} = \begin{cases} a_{ij} & \text{when } i \neq j \\ 0 & \text{when } i = j. \end{cases}$$

We then obtain from "truncated" backward equation (8):

$$P_{ij}(t,\ n) = \delta_{ij} + \int_0^t \sum_{k=0}^{n} a_{ik} P_{kj}(u,\ n)\, du = \delta_{ij} + a_{ii} \int_0^t P_{ij}(u,\ n)\, du$$

$$+ \int_0^t \Big(\sum_{k \geqslant 0} a_{ik} P_{kj}(u,\ n) \Big)\, du.$$

Putting $c_k(n) = \alpha_{ik} P_{kj}(u,\ n)$, we derive from the lemma:

$$\sum_{k \geqslant 0} \alpha_{ik} P_{kj}(u,\ n) \to \sum_{k \geqslant 0} \alpha_{ik} Q_{kj}(u).$$

By the Beppo Levi theorem, we can pass to the limit under the integral sign, which gives (12).

We now verify equality (13). It follows from (8) and (10) that

$$\sum_k a_{ik} P_{kj}(t,\ n) = \sum_k P_{ik}(t,\ n)\, \alpha_{kj},$$

or

$$a_{ii} P_{ij}(t,\ n) + \sum_{k \geqslant 0} \alpha_{ik} P_{kj}(t,\ n) = \sum_{k \geqslant 0} P_{ik}(t,\ n)\, \alpha_{kj} + P_{ij}(t,\ n)\, a_{jj}.$$

In view of the lemma, we can pass to the limit under the summation sign, which gives (13).

3. Putting $c_k(n) = P_{ik}(s,\ n)\, P_{kj}(t,\ n)$ in the equality

$$P_{ij}(s+t,\ n) = \sum_k P_{ik}(s,\ n)\, P_{kj}(t,\ n),$$

we obtain from the lemma:

$$Q_{ij}(s+t) = \sum_k Q_{ik}(s)\, Q_{kj}(t).$$

Proof of Proposition 4. A. We first prove the following **minimal property** of the above solution $Q(t)$ to the Kolmogorov forward and backward equation, viz., if $P(t)$, $t \geqslant 0$, is a semi-stochastic matrix function and a solution of the Kolmogorov forward and/or backward equation, then $P_{ij}(t) \geqslant Q_{ij}(t)\ \forall i,\ j,\ t$.

In fact, let, $P(t)$, $t \geqslant 0$, satisfy the Kolmogorov backward equation

$$P'(t) = A \cdot P(t), \quad P(0) = I.$$

In particular, putting

$$x_i(t) = P_{ij}(t),\ x_i(0) = \delta_{ij}$$

for a certain j, we have

$$x_i'(t) = \sum_k a_{ik} x_k(t),$$

whence, denoting

$$x(t,\ n) = \{x_0(t,\ n),\ x_1(t,\ n),\ \ldots,\ x_n(t,\ n)\}^T,\ x_i(t,\ n) = x_i(t),\ i \leqslant n,$$

we get

$$x_i'(t,\ n) = \sum_{k=0}^{n} a_{ik} \cdot x_k(t,\ n) + \sum_{k>n} a_{ik} x_k(t)$$

$$= \sum_{k=0}^{n} a_{ik}(n) \cdot x_k(t,\ n) + b_i(t,\ n),$$

or, in vector form,

$$x'(t,\ n) = A(n)\, x(t,\ n) + b(t,\ n).$$

The solution of this differential equation is of the form

$$x_i(t,\ n) = P(t,\ n) \cdot x(0,\ n) + \int_0^t P(t-u,\ n)\, b(u,\ n)\, du,$$

or

$$x_i(t,\ n) = P_{ij}(t,\ n) + \int_0^t \sum_{k=0}^{n} P_{ik}(t-u,\ n)\, b_k(u,\ n)\, du, \quad i \leqslant n, \quad j \leqslant n,$$

whence, due to $b_i(u,\ n) \geqslant 0$, and $x_i(t,\ n) = P_{ij}(t)$, we have

$$P_{ij}(t) \geqslant P_{ij}(t,\ n),$$

which gives

$$P_{ij}(t) \geqslant Q_{ij}(t)$$

as $n \to \infty$.

B. Now, let the Kolmogorov backward equation have a unique solution in the class of semi-stochastic matrix functions, being itself a stochastic matrix function. By Proposition 3, such a solution is

$Q(t)$, while

$$\sum_j Q_{ij}(t) = 1 \quad \forall\, i.$$

By Proposition 3 again, the forward equation also has a solution in the class of semi-stochastic matrix functions. If $P(t)$ is one of such solutions, then, due to the minimal property (see Item A), we must have $P_{ij}(t) \geqslant Q_{ij}(t)\ \forall\ i, j$, which, combined with the inequality

$$\sum_j P_{ij}(t) \leqslant 1,$$

yields $P_{ij}(t) = Q_{ij}(t)$, i.e., a solution of the forward equation is a unique stochastic matrix function.

The converse statement of Proposition 4 is proved similarly.

Chapter 4

Renewal Processes

§ 1. Definition of Renewal Process

Let $\{z_k : k \geqslant 1\}$ be a sequence of nonnegative RRV's. Put

$$t_n = z_1 + \ldots + z_n, \quad n \geqslant 1, \ t_0 = 0,$$
$$\nu(t) = \sup\{n : t_n < t\}, \ t \geqslant 0.$$

E.g., if the variables z_k are understood as the durations of existence of a sequence of replaceable elements, then t_n is the epoch of changing the n-th element, and $\nu(t)$ the number of elements replaced up to t.

The process $\nu(t)$ is called a **renewal process**. Thus, any sequence $\{z_k : k = 1, 2, \ldots\}$ of real positive semi-definite RV's determines a renewal process $\nu(t)$, due to which a renewal process is said to be **a sequence** $\{z_k\}$ of nonnegative RRV's. A renewal process is regarded as given if, for each integer $n \geqslant 1$, a distribution of the random vector $(z_1, \ldots, z_n)$ is given.

A renewal process $\{z_k : k \geqslant 1\}$ is said to be **delayed recurrent** if the RV's $z_1, z_2, z_3, \ldots$ are i.i.d., beginning with the second.

However, in the case where all the RV's $z_1, z_2, \ldots$ are distributed identically (and are independent), we will speak of a **recurrent renewal process.**

Below, we only consider delayed recurrent renewal processes. Put

$$A_1(t) = P(z_1 < t), \ A(t) = P(z_k < t), \ k \geqslant 2.$$

In the case of a recurrent renewal process,

$$A_1(t) = A(t).$$

Assume additionally that $P(z_k = 0) < 1$ for all $k \geqslant 1$.

Proposition. There exists a number $\theta_0 > 0$ such that

$$E\{\exp(\theta\nu(t))\} < \infty$$

for any $t \geqslant 0$ and any $\theta \leqslant \theta_0$. In particular, for any $t \geqslant 0$, the random number $\nu(t)$ has finite moments of all orders (since $x^n \leqslant n!\, e^x\ \forall\ x \geqslant 0$).

Proof. It suffices to prove the statement only for the case of a recurrent renewal process. Since $P(z_k = 0) < 1$, there exists a number $\delta > 0$ such that $P(z_k \geqslant \delta) = \varepsilon > 0$. Define a new recurrent renewal process $\{z_k' : k \geqslant 1\}$ by the equalities

$$z_k' = \begin{cases} \delta & \text{if } z_k \geqslant \delta \\ 0 & \text{if } z_k < \delta \end{cases}.$$

Since $z_k' \leqslant z_k$, $\nu'(t) \geqslant \nu(t)$. But the RV $\nu'(t) - \left[\frac{t}{\delta}\right]$ has the negative binomial distribution, which entails the formulated statement.

For a delayed recurrent renewal process, we put

$$H_1(t) = E\nu(t).$$

In the case $A_1(t) = A(t)$, we will denote

$$H(t) = E\nu(t).$$

Since

$$\nu(t) = \sum_{k \geqslant 1} 1_{\{t_k < t\}}, \quad E1_{\{t_k < t\}} = P(t_k < t) = A_k(t),$$
$$A_{k+1}(t) = A_k(t) * A(t) = A_1(t) * P(z_2 + \ldots + z_{k+1} < t),$$
$$E\nu(t) = \sum_{k \geqslant 1} E1_{\{t_k < t\}} = \sum_{k \geqslant 1} A_k(t),$$

we have

$$H_1(t) = A_1(t) + \int_0^t H_1(t-u)\, dA(u), \tag{1}$$

$$H(t) = A(t) + \int_0^t H(t-u)\, dA(u), \tag{2}$$

and

$$H_1(t) = A_1(t) + \int_0^t A_1(t-u)\, dH(u). \tag{3}$$

Equation (2) is called a **renewal equation**, and the function H_1 or H a **renewal function.**

§ 2. Elementary Renewal Theorem

The basic results of the theory of renewal processes are represented as three theorems, viz., the elementary renewal theorem (proved in this section), Blackwell theorem (see § 3), and key renewal theorem (see § 4).

Proposition (Elementary Renewal Theorem).

$$\lim_{t\to+\infty}\frac{H_1(t)}{t}=a, \text{ where } \int_0^\infty t\,dA(t)=a^{-1},$$

including the case $\int_0^\infty t\,dA(t)=+\infty$, and assuming that $a=0$. The number a is called **renewal intensity.**

Proof. 1°. We first show that if there exists one of the limits

$$\lim_{t\to\infty}\frac{H(t)}{t} \text{ or } \lim_{t\to\infty}\frac{H_1(t)}{t}, \tag{1}$$

then there also exists the other, and

$$\lim_{t\to\infty}\frac{H(t)}{t}=\lim_{t\to\infty}\frac{H_1(t)}{t}. \tag{2}$$

Indeed, for $0<\varepsilon<1$, we have from (3), § 1:

$$H_1(t)\leqslant 1+H(T),$$

$$H_1(t)\geqslant\int_0^{T-\varepsilon T}A_1(T-u)\,dH(u)\geqslant A_1(\varepsilon T)\,H(T-\varepsilon T),$$

or

$$\frac{H_1(T)}{T}\leqslant\frac{H(T)}{T}+\frac{1}{T},$$

$$\frac{H_1(T)}{T}\geqslant\frac{H(T-\varepsilon T)}{T-\varepsilon T}(1-\varepsilon)\,A_1(\varepsilon T).$$

Due to the arbitrariness of $\varepsilon>0$, it follows from the latter inequalities that if there exists one of limits (1), then the other also exists, and (2) is fulfilled.

2°. Consider the case $a^{-1}=a\int_0^\infty t\,dA(t)<\infty$, and take $A_1(t)$ in the special form

$$A_1(t)=a\int_0^\infty[1-A(u)]\,du,\ t\geqslant 0. \tag{3}$$

A_1 is a DF, since $a\int_0^\infty[1-A(u)]\,du=1$, and the function A_1 is non-decreasing, continuous, and $A(+\infty)=1$ (assuming that $A_1(t)=0$ for $t<0$).

In the Laplace-Stieltjes transformations, (1), § 1, and (3) will assume the form

$$h_1(s)=\alpha_1(s)+h_1(s)\,\alpha(s),$$

$$\alpha_1(s)=\frac{a}{s}[1-\alpha(s)],$$

whence

$$h_1(s)=\frac{a}{s},$$

i.e.,

$$H_1(t)=at.$$

Therefore, using the statement of Item 1°, we conclude that, first, $\lim_{t\to\infty}\frac{H(t)}{t}$ exists and equals a, and, second, there exists $\lim_{t\to\infty}\frac{H_1(t)}{t}$ equal to a for an arbitrary DF A_1.

3°. It remains to investigate the case $\int_0^\infty t\,dA(t)=\infty$. For an arbitrary $\varepsilon>0$, we also consider a delayed recurrent renewal process $\{z_i'^{k}\}$ by putting

$$z_k'=\min(z_k,\ \varepsilon^{-1}),$$

whence

$$0\leqslant\lim_{t\to\infty}\frac{H_1(t)}{t}\leqslant\lim_{t\to\infty}\frac{H_1'(t)}{t}=a_\varepsilon=1/Ez_k'\to 0 \quad \text{for } \varepsilon\downarrow 0,$$

which was required to prove.

§ 3. Blackwell Theorem

A point x is called a **point of increase** of a DF A if, for any two numbers a and b such that $a<x<b$, the inequality $A(b)-A(a)>0$ is true. A distribution A is said to be **arithmetic** (or **lattice**) if there exists a number $\lambda>0$ such that any point of increase of the DF A is a multiple of λ, i.e., of the form $n\lambda$, where n is an integer. If there exists no such λ, then the distribution A is said to be **non-arithmetic**. Note that an arithmetic distribution corresponds to the distribution of an RV such that assumes (with probability 1) only values of the form $n\lambda$, where $\lambda>0$, and n is an integer. We give the following statement without proof.

Proposition (Blackwell Theorem). If A is a non-arithmetic distribution, then, for any number h,

$$H(t+h) - H(t) \to ah \tag{1}$$

as $t \to +\infty$.

Proof can be seen in *An Introduction to Probability Theory and Its Applications* by W. Feller (John Wiley & Sons, Inc., New York-London-Sydney, 1968), V. 2, Ch. 11.

§ 4. Key Renewal Theorem

We first introduce the concept of directly Riemann integrable function on $[0, \infty)$. Let Q be a Borel measurable numerical function defined on $[0, \infty)$. For a number $h > 0$, we denote the lower and upper bounds of the function Q in the interval $(k-1)h \leqslant x < kh$, $k \geqslant 1$, by m_k and M_k, respectively. Assume that the two series

$$s = h \sum_{k \geqslant 1} m_k \quad \text{and} \quad S = h \sum_{k \geqslant 1} M_k$$

converge absolutely. We say that the function Q is **directly Riemann integrable** on $[0, \infty)$ if $S - s \to 0$ when $h \downarrow 0$. In this case, we put

$$\int_0^\infty Q(t)\,dt = \lim_{h \downarrow 0} s = \lim_{h \downarrow 0} S. \tag{1}$$

Proposition (Key Renewal Theorem). Let Q be a directly Riemann integrable function on $[0, \infty)$. If A is a non-arithmetic distribution, then

$$\lim_{t \to \infty} \int_0^t Q(t-x)\,dH(x) = a \int_0^\infty Q(t)\,dt \tag{2}$$

(the integral on the left is understood in the sense of Lebesgue-Stieltjes: it exists, since Q is a bounded and Borel measurable function, whereas H monotonic, and the integral on the right the usual Riemann integral).

Remark. Recall that the usual Riemann integral on $[0, \infty)$ is defined as the limit of a Riemann integral on $[0, a]$ when $a \to +\infty$. Direct integrability coincides with the usual Riemann integrability for a function Q equal to zero outside a certain finite interval. The same is true with monotonic functions Q. However, it is not complicated to give an example of a continuous Riemann integrable function on $[0, \infty)$, which is not directly integrable. On the other hand, any directly Riemann integrable function on $[0, \infty)$ is Riemann integrable in the usual sense, which follows from the inequality

$$\left| \int_0^a Q(t)\,dt - \sigma \right| \leqslant |S-s| + |s-\sigma| + h \sum_{kh > a} m_k,$$

where $a > 0$, $\sigma = \lim_{h \downarrow 0} s$ and from the usual Riemann integrability of the function Q on any finite interval.

Proof is broken into several parts.

1°. For $h > 0$, denote by $w_n(t)$ a function assuming the value 1 when $(n-1)h \leqslant t < nh$, and 0 outside their interval. Put

$$m(t) = \sum_{n \geqslant 1} m_n w_n(t), \quad M(t) = \sum_{n \geqslant 1} M_n w_n(t).$$

Then, for all $t \geqslant 0$,

$$m(t) \leqslant Q(t) \leqslant M(t);$$

in particular,

$$\int_0^t m(t-u)\,dH(u) \leqslant \int_0^t Q(t-u)\,dH(u) \leqslant \int_0^t M(t-u)\,dH(u).$$

Now, if

$$\lim_{t \to \infty} \int_0^t m(t-u)\,dH(u) = as, \quad \lim_{t \to \infty} \int_0^t M(t-u)\,dH(u) = aS, \tag{3}$$

then

$$as \leqslant \varliminf_{t \to \infty} \int_0^t Q(t-u)\,dH(u) \leqslant \varlimsup_{t \to \infty} \int_0^t Q(t-u)\,du \leqslant aS,$$

in which case (2) holds because of (1).

2°. Thus, it suffices to prove (3), for which we shall first verify that there exists a number $C = C(h)$ such that

$$0 \leqslant H(t+h) - H(t) \leqslant C \tag{4}$$

for all numbers t. To this end, writing formula (2), § 1, in the form

$$H(t) = A(t) + \int_0^t A(t-u)\,dH(u),$$

we obtain

$$H(t+h) \leqslant 1 + \int_0^t A(t+h-u)\,dH(u) + \int_t^{t+h} A(t+h-u)\,dH(u)$$
$$\leqslant 1 + H(t) + A(h)\,[H(t+h) - H(t)],$$

or

$$[H(t+h) - H(t)]\,[1 - A(h)] \leqslant 1. \tag{5}$$

If $1 - A(h) > 0$, then it suffices to take $(1 - A(h))^{-1}$ as $C(h)$. If, however, $1 - A(h) = 0$, then we shall make use of the fact that there exists a number $\varepsilon > 0$ such that $1 - A(\varepsilon) > 0$; hence, due

to (5), for any integer $m \geqslant 1$,

$$H(t+m\varepsilon)-H(t)=\sum_{k=0}^{m-1}[H(t+k\varepsilon+\varepsilon)-H(t+k\varepsilon)]\leqslant\frac{m}{1-A(\varepsilon)};$$

therefore, when $0\leqslant h\leqslant m\varepsilon$,

$$0\leqslant H(t+h)-H(t)\leqslant H(t+m\varepsilon)-H(t)\leqslant\frac{m}{1-A(\varepsilon)}=C_1(h).$$

3°. We can now prove (3). Since

$$\int_0^t w_n(t-u)\,dH(u)=H(t-nh+h)-H(t-nh),$$

we obtain

$$\int_0^t m(t-u)\,dH(u)=\sum_{n\geqslant 1} m_n[H(t-nh+h)-H(t-nh)].$$

Putting

$$w(t)=\int_0^t m(t-u)\,dH(u),$$

$$J_N(t)=\sum_{n=1}^{N} m_n[H(t-nh+h)-H(t-nh)],$$

we get, taking into account (4):

$$J_N(t)-C\sum_{n>N}|m_n|\leqslant w(t)\leqslant J_N(t)+C\sum_{n>N}|m_n|.$$

By the Blackwell theorem (§ 3),

$$\lim_{t\to\infty} J_N(t)=ah\sum_{n=1}^{N} m_n;$$

therefore,

$$ah\sum_{n=1}^{N} m_n-C\sum_{n>N}|m_n|$$

$$\leqslant\underline{\lim_{t\to\infty}}\, w(t)\leqslant\overline{\lim_{t\to\infty}}\, w(t)\leqslant ah\sum_{n=1}^{N} m_n+C\sum_{n>N}|m_n|,$$

and, since the series $\sum_{n\geqslant 1} m_n$ converges absolutely, while the number $N \geqslant 1$ can be selected arbitrarily, this implies that the first part of formula (3) is valid. The second part is proved similarly.

Corollary. Let

(1) Q be a directly Riemann integrable function on $[0, \infty)$,

(2) A a non-arithmetic distribution on $[0, \infty)$,

and

(3) W a solution of the equation

$$W = Q + A * W. \tag{6}$$

Then

$$\lim_{t\to\infty} W(t)=a\int_0^\infty Q(x)\,dx, \tag{7}$$

where

$$a^{-1}=\int_0^\infty x\,dA(x)=\int_0^\infty [1-A(x)]\,dx.$$

Proof follows from the representation

$$W = Q + Q * A + Q * A^{*2} + \ldots = Q + Q * H,$$

formula (2), and the equality

$$\lim_{t\to\infty} Q(t) = 0.$$

Example 1. Let Q be a Borel measurable and Lebesgue integrable function on $[0, \infty)$. Show that

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T\left\{\int_0^t Q(t-u)\,dH_1(u)\right\}dt=a\int_0^\infty Q(t)\,dt.$$

Hint: By changing the order of integration, verify that

$$\frac{1}{T}\int_0^T\int_0^t Q(t-u)\,dH_1(u)\,dt=\frac{H_1(T)}{T}\int_0^T Q(v)\,dv+\varepsilon_T,$$

$$|\varepsilon_T|\leqslant\frac{H_1(T)}{T}\int_0^T|Q(v)|\,dv-\frac{H_1(T-u_0)}{T}\int_0^{u_0}|Q(v)|\,dv,\quad 0\leqslant u_0\leqslant T.$$

Example 2. Prove that, for $h \geqslant 0$,

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T[H_1(t+h)-H_1(t)]\,dt=ah.$$

Hint: In the previous problem, put

$$Q(t)=\begin{cases}1 \text{ when } 0\leqslant t<h,\\ 0 \text{ when } t\geqslant h,\end{cases}$$

Example 3. Let Q be a directly Riemann integrable function on $[0, \infty)$, and A a non-arithmetic distribution. Then

$$\lim_{t\to\infty} \int_0^t Q(t-u)\, dH_1(u) = a \int_0^\infty Q(t)\, dt.$$

Hint: In view of (1), § 1,

$$Q * H_1 = W = Q * A_1 + A * W.$$

Then use the corollary, taking into account that

$$\int_0^\infty (Q * A_1)(t)\, dt = \int_0^\infty \left\{ \int_u^\infty Q(t-u)\, dt \right\} dA_1(u)$$

$$= \int_0^\infty dA_1(u) \cdot \int_0^\infty Q(t)\, dt = \int_0^\infty Q(t)\, dt.$$

Example 4. If A is a non-arithmetic distribution, then

$$H_1(t+h) - H_1(t) \to ah \text{ as } t \to +\infty.$$

Hint: See Example 2.

Chapter 5

Regenerative Processes

§ 1. Definition of Regenerative Process

We call an ordered pair (z, ξ) a **cycle** of duration z if z is a non-negative real (random) number, while ξ a random function defined on $[0, \infty)$, and assuming values in a measurable space $(X, \mathbf{B})$. It will always be assumed that the function $(t, w) \to \xi(t, w)$ is measurable, whereas

$$P(z = 0) < 1, \quad P(z < +\infty) = 1.$$

Consider a sequence $\{(z_k, \xi_k) : k \geqslant 1\}$ of cycles, and assume that its cycles are stochastically independent (each cycle is regarded as a collection of RV's, and, beginning with the second, stochastically equivalent. In particular, the sequence $\{z_k : k \geqslant 1\}$ of randon numbers forms a delayed recurrent renewal process for which we put

$$A_1(x) = P(z_1 < x), \quad A(x) = P(z_k < x), \quad k \geqslant 2.$$

Now, we construct a process $\xi(t)$, $0 \leqslant t < \infty$, with values in $(X, \mathbf{B})$ by putting

$$\xi(t) = \begin{cases} \xi_1(t) & \text{when } 0 \leqslant t < z_1 \\ \xi_2(t - z_1) & \text{when } z_1 \leqslant t < z_1 + z_2, \\ \dots\dots\dots\dots\dots\dots\dots\dots & \\ \xi_k(t - t_{k-1}) & \text{when } t_{k-1} \leqslant t < t_k \\ \dots\dots\dots\dots\dots\dots\dots\dots & \end{cases}$$

where

$$t_0 = 0, \quad t_k = z_1 + \dots + z_k, \quad k \geqslant 1.$$

The process $\xi(t)$, $t \geqslant 0$, so defined is called a **regenerative process**, and the epochs $t_1, t_2, \dots$ **regeneration epochs**.

Let $B \in \mathbf{B}$. The principal goal of the theory of regenerative processes is to indicate the conditions under which the limit

$$\lim_{t\to\infty} P\{\xi(t) = B\} \tag{1}$$

exists, and also a method of its computation.

Put

$$\mu'_B(t) = P\{\xi_1(t) \in B, \ z_1 > t\} = P\{\xi(t) \in B, \ z_1 > t\},$$

$$\mu_B(t) = P\{\xi_k(t) \in B, \ z_k > t\} = P\{\xi(t_{k-1} + t) \in B, \ z_k > t\}, \quad k \geqslant 2.$$

The main technique for calculating limit (1) is based on the relation

$$P\{\xi(t) \in B\} = \mu'_B(t) + \sum_{k\geqslant 1} \int_0^t \mu_B(t-x)\, dA_k(x),$$

or

$$P\{\xi(t) \in B\} = \mu'_B(t) + \int_0^t \mu_B(t-x)\, dH_1(x) \tag{2}$$

which is obtained from the total probability formula. Meanwhile,

$$A_k(x) = P(t_k < x), \quad H_1(x) = \sum_{k\geqslant 1} A_k(x)$$

is the renewal function for the renewal process $\{z_k : k \geqslant 1\}$.

In the subsequent section, we shall see how the key renewal theorem (more precisely, Example 3, § 4, Ch. 4) will permit us to find (1) by means of relation (2).

§ 2. Limit Theorem for Regenerative Process

Proposition. Let

(1) A be a non-arithmetic distribution,

and

(2) there exist an integer $n \geqslant 0$ such that the function $Q(t) = \int_0^t \mu_B(t-x)\,dF(x)$, $t \geqslant 0$, where $F(x) = A^{*n}(x)$, is directly Riemann integrable on $[0, \infty)$.

Then

$$\lim_{t\to\infty} P\{\xi(t) \in B\} = a \int_0^\infty \mu_B(x)\,dx, \quad a^{-1} = \int_0^\infty x\,dA(x).$$

Proof. If $n = 0$, then $Q(t) = \mu_B(t)$, and the statement follows from the key renewal theorem (see Example 3, § 4, Ch. 4) as well as from the fact that

$$0 \leqslant \mu_B'(t) \leqslant 1 - A_1(t) \to 0 \text{ when } t \to \infty.$$

The same argument is also applicable for $n \geqslant 1$, since it then follows from (2), § 1, that

$$P\{\xi(t) \in B\} = \mu_B'(t) + \sum_{k=1}^{n} \int_0^t \mu_B(t-x)\,dA_k(x) + \int_0^t Q(t-x)\,dH_1(x);$$

besides, as $t \to +\infty$

$$0 \leqslant \int_0^t \mu_B(t-x)\,dA_k(x) \leqslant \int_0^t [1 - A(t-x)]\,dA_k(x)$$
$$= A_k(t) - A_{k+1}(t) \to 0.$$

Corollary. Let A be a non-arithmetic distribution, and at least one of the following conditions fulfilled.

C1. The function $\mu_B(t)$ is Riemann integrable on each finite interval, and $a^{-1} = \int_0^\infty x\,dA(x) < +\infty$.

C2. For a certain integer $n \geqslant 1$, the function $A^{(n)}$ defined by the equalities

$$A^{(1)} = A, \quad A^{(k+1)}(t) = \int_0^t A^{(k)}(t-x)\,dA(x), \quad k \geqslant 1,$$

is absolutely continuous, and $\int_0^\infty x\,dA(x) < \infty$.

Then

$$\lim_{t\to\infty} P\{\xi(t) \in B\} = a \int_0^\infty \mu_B(x)\,dx.$$

Proof. 1°. Let C1 be fulfilled. Then the function $Q = \mu_B$ satisfies the following conditions:

(1) it is Borel measurable,

(2) it is Riemann integrable in the usual sense on $[0, \infty)$, and

(3) $0 \leqslant Q(t) \leqslant G(t) = 1 - A(t)$, where the function $G(t)$ is directly Riemann integrable on $[0, \infty)$.

It follows from the same conditions that the function Q is directly Riemann integrable (show!).

2°. Now, let C2 be fulfilled. Put

$$Q(t) = \int_0^t \mu_B(t-x)\,dA^{(n)}(x), \quad t \geqslant 0.$$

It will be shown below that the function Q is continuous; besides,

$$0 \leqslant Q(t) \leqslant \int_0^t [1 - A(t-x)]\,dA^{(n)}(x) = A^{(n)}(t) - A^{(n+1)}(t) = G(t),$$

and the function G is directly Riemann integrable on $[0, \infty)$, since

$$G(t) = [1 - A^{(n+1)}(t)] - [1 - A^{(n)}(t)],$$
$$\int_0^\infty [1 - A^{(k)}(t)]\,dt = \int_0^\infty t\,dA^{(k)}(t) = k \int_0^\infty t\,dA(t) < +\infty,$$

whereas each of the functions $1 - A^{(k)}(t)$ is monotonic. Therefore, C1-C3 of the previous item are fulfilled, and the function Q is directly Riemann integrable on $[0, \infty)$. It now remains to apply the statement proved above.

To prove the continuity of the function Q, we make use of the Radon and Lusin theorems. Denote by P the probability measure generated by the distribution $A^{(n)}$. Since the measure P is absolutely continuous (with respect to the Lebesgue linear measure on the straight line), there exists such an integrable function p on $[0, \infty)$ that

$$P(M) = \int_M p(x)\,dx$$

for any measurable set $M \subset [0, \infty)$. We have

$$Q(t) = \int_0^t \mu_B(t-x)\,p(x)\,dx.$$

Since, for $h \geqslant 0$,

$$Q(t+h) = \int_0^{t+h} \mu_B(t+h-x)\, p(x)\, dx = \int_{-h}^{t} \mu_B(t-u)\, p(u+h)\, du,$$

we obtain

$$Q(t+h) - Q(t) = \int_{-h}^{0} \mu_B(t-u)\, p(u+h)\, du + \int_0^t \mu_B(t-u)\, [p(u+h) - p(u)]\, du;$$

therefore, due to $0 \leqslant \mu_B(t) \leqslant 1$,

$$|Q(t+h) - Q(t)| \leqslant A^{(n)}(h) + \int_0^t |p(u+h) - p(u)|\, du.$$

Similarly,

$$|Q(t-h) - Q(t)| \leqslant \int_0^t |p(u+h) - p(u)|\, du.$$

Because $A^{(n)}(h) \to 0$ as $h \downarrow 0$, it siffices to show that

$$\int_0^t |p(u+h) - p(u)|\, du \to 0 \text{ when } h \to 0,$$

or

$$\int_0^t |p_n(u+h) - p_n(u)|\, du \to 0 \text{ when } h \to 0, \tag{1}$$

with

$$p_n(u) = \min\{n,\ p(u)\},$$

which is due to the fact that, for any $t \geqslant 0$,

$$\int_0^t |p(u) - p_n(u)|\, du = \int_0^t p(u)\, du - \int_0^t p_n(u)\, du \to 0$$

as $n \to +\infty$.

To prove (1), we use the Lusin theorem, according to which, for a chosen $T > 0$ and any $\varepsilon > 0$, there exists such a continuous function φ on $[0,\ T]$ that the Lebesgue measure of the set

$$M = \{x \in [0,\ T] : p_n(x) \neq \varphi(x)\}$$

is less than ε, and $|\varphi(x)| \leqslant n$. Let $T > t+h$. Then

$$\begin{aligned}
\int_0^t |p_n(u+h) - p_n(u)|\, du &\leqslant \int_0^t |\varphi(u+h) - \varphi(u)|\, du \\
&\quad + \int_0^t |\varphi(u) - p_n(u)|\, du \\
&\quad + \int_0^t |\varphi(u+h) - p_n(u+h)|\, du \\
&\leqslant \int_0^t |\varphi(u+h) - \varphi(u)|\, du \\
&\quad + 2\int_M |\varphi(u) - p_n(u)|\, du \\
&\leqslant \int_0^t |\varphi(u+h) - \varphi(u)|\, du + 4n\varepsilon,
\end{aligned}$$

and the value of the latter integral tends to zero as $h \to 0$ due to the uniform convergence of the function φ on $[0,\ T]$.

Example 1. Let

(1) A be a non-arithmetic distribution, and

(2) there exist an integer $n \geqslant 0$ such that the function

$$Q(t) = \int_0^t \mu_B(t-x)\, dA_n(x)$$

is directly Riemann integrable on $[0,\ \infty)$.

Show that

$$\lim_{t\to\infty} P\{\xi(t) \in B\} = a \int_0^\infty \mu_B(x)\, dx.$$

Example 2. Let

(1) $a^{-1} = \int_0^\infty x\, dA(x) < \infty$,

(2) $\mu_{B'}(t)$ a Lebesgue measurable function, whereas

(3) $\mu_B(t)$ a Borel measurable function (and, therefore, Lebesgue integrable on $[0,\ \infty)$, since $0 \leqslant \mu_B(t) \leqslant 1 - A(t)$, $\int_0^\infty [1 - A(t)]\, dt = a^{-1} < \infty$).

Then

(1) $P\{\xi(t) \in B\}$ is a Lebesgue integrable function on any interval $[0, T]$,

(2) there exists

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T P\{\xi(t) \in B\}\, dt = P_B,$$

and

$$P_B = a \int_0^\infty \mu_B(t)\, dt. \tag{3}$$

Hint: See example 1, § 4, Ch. 4, and formula (2), § 1.

§ 3. Law of Large Numbers for Regenerative Processes

Let $\xi(t)$ be a regenerative process with values in a certain measurable space $(X, \mathbf{B})$ for each $t \geqslant 0$, $t_1, t_2, \ldots$ consecutive regeneration epochs of the process, and $z_k = t_k - t_{k-1}$, $k \geqslant 1$, $t_0 = 0$. Then the sequence of RV's $\{z_k, k \geqslant 1\}$ forms a delayed recurrent renewal process. Put $A_1(x) = P(z_1 < x)$, $A(x) = P(z_k < x)$ for $k > 1$. Assume that

$$0 < a^{-1} = Ez_k = \int_0^\infty x\, dA(x) < \infty, \quad k > 1,$$

and, for a certain set $B \in \mathbf{B}$, put for $k \geqslant 1$, $t \geqslant 0$

$$\mu_1(t) = P\{\xi(t) \in B,\ z_1 > t\},\ \mu(t) = P\{\xi(t_k + t) \in B,\ z_{k+1} > t\}.$$

Finally, denote the indicator of the event $\{\xi(t) \in B\}$ by $x(t)$.

Proposition. If each trajectory of the process $\{x(t)\}$, and the function μ_1 are Lebesgue measurable, whereas μ is Borel measurable, then the function $P(t) = P\{\xi(t) \in B\}$ is Lebesgue measurable, while, for any $\lambda > 0$,

$$E\left|\frac{1}{T}\int_0^T x(t)\, dt - p\right|^\lambda \to 0, \tag{1}$$

$$\left|\frac{1}{T}\int_0^T P(t)\, dt - p\right| \to 0, \quad p = a\int_0^\infty \mu(t)\, dt \tag{2}$$

as $T \to \infty$.

Remark. $\bar{x}(T) = \frac{1}{T}\int_0^T x(t)\, dt$ is the sojourn of the process $\xi(t)$ in the set B in time T.

Proof. 1°. As regards (2), see Example 2, § 2. It suffices to verify statement (1) for $\lambda = 2$, since $0 \leqslant \bar{x}(T) \leqslant 1$, $0 \leqslant p \leqslant 1$; therefore $|\bar{x}(T) - p| \leqslant 1$, and

$$E|\bar{x}(T) - p|^\lambda \leqslant \begin{cases} E|\bar{x}(T) - p|^2 & \text{when } \lambda \geqslant 2 \\ E^{\lambda/2}|\bar{x}(T) - p|^2 & \text{when } 0 \leqslant \lambda \leqslant 2. \end{cases}$$

2°. If $\lambda = 2$, the left-hand side of (1) is written in the form

$$E[\bar{x}(T) - p]^2 = E\bar{x}^2(T) - p^2 - 2p[E\bar{x}(T) - p].$$

According to (2),

$$E\bar{x}(T) - p = \frac{1}{T}\int_0^T P(t)\, dt - p \to 0$$

as $T \to \infty$.

It remains to prove that

$$\overline{\lim_{T\to\infty}} E\bar{x}^2(T) \leqslant p^2. \tag{3}$$

3°. Note that

$$E\bar{x}^2(T) = \frac{1}{T^2}\int_0^T\int_0^T Ex(s)\, x(\tau)\, ds d\tau = \frac{2}{T^2}\iint_{s+t\leqslant T} Ex(s)\, x(s+t)\, ds dt \tag{4}$$

$$= \frac{2}{T^2}\iint_{s+t\leqslant T} P\{\xi(s) \in B,\ \xi(s+t) \in B\}\, ds dt,$$

where $s \geqslant 0$, $t \geqslant 0$. When $s > 0$, $t > 0$, the event $\{\xi(s) \in B, \xi(s+t) \in B\}$ is contained in the union of the following events, viz.,

E_1 "the points s and $s + t$ get in the same regeneration cycle",

E_2 "the point s gets in the first regeneration cycle", and

E_0 "the points s and $s + t$ get in different regeneration cycles not identical with the first, and $\xi(s) \in B$, $\xi(s+t) \in B$".

Therefore

$$P\{\xi(s) \in B,\ \xi(s+t) \in B\} \leqslant P(E_1) + P(E_2) + P(E_0). \tag{5}$$

4°. We have

$$P(E_1) = \sum_{k\geqslant 0} P(t_k < s,\ t_{k+1} \geqslant s+t) = \sum_{k\geqslant 0} P(s+t-z_{k+1} \leqslant t_k < s) \tag{6}$$

$$= 1 - A_1(s+t) + \sum_{k\geqslant 1}\int_0^s [1 - A(s+t-x)]\, dA_k(x)$$

$$= 1 - A_1(s+t) + \int_0^s [1 - A(s+t-x)]\, dH_1(x),$$

where

$$A_k(x) = P(t_k < x),\ H_1(x) = \sum_{k\geqslant 1} A_k(x);$$

(7) $$P(E_2)=P(z_1\geqslant s)=1-A_1(s),$$

$$P(E_0)=\sum_{n\geqslant 1}\sum_{k\geqslant 1}P\{\xi(s)\in B,\ t_n<s,\ t_{n+1}\geqslant s,$$

$$\xi(s+t)\in B,\ t_{n+k}<s+t,\ t_{n+k+1}\geqslant s+t\}$$

$$=\sum_{n\geqslant 1}\sum_{k\geqslant 1}\int_0^s \mu(s-u)\,dA_n(u)\int_s^{s+t}\mu(s+t-v)\,dA_{n+k}(v)$$

$$=\int_0^s \mu(s-u)\,dH_1(u)\int_s^{s+t}\mu(s+t-v)\,d_vH(v-u),$$

where $H(x)=\sum_{k\geqslant 1}A^{*k}(x)$, and A^{*k} is the k-th Stieltjes convolution of the DF A; besides, we have used the equalities

$$A_{n+k}(v)=P(t_{n+k}<v\mid t_n=u)$$
$$=P(z_{n+1}+\ldots+z_{n+k}<v-u)=A^{*k}(v-u)$$

for $t_n=u$.

5°. Show that

(9) $$I_1=\frac{1}{T^2}\iint_{s+t\leqslant T}[1-A_1(s+t)]\,dsdt$$
$$=\frac{1}{T^2}\int_0^T u\,[1-A_1(u)]\,du,$$

(10) $$I_2=\frac{1}{T^2}\iint_{s+t\leqslant T}dsdt\int_0^s[1-A(s+t-x)]\,dH_1(x)$$
$$\leqslant\frac{H_1(T)}{T}\cdot\frac{1}{T}\int_0^T u\,[1-A(u)]\,du,$$

(11) $$I_3=\frac{1}{T^2}\iint_{s+t\leqslant T}[1-A_1(s)]\,dsdt$$
$$=\frac{1}{T}\int_0^T[1-A_1(s)]\,ds-\frac{1}{T^2}\int_0^T s\,[1-A_1(s)]\,ds,$$

(12) $$I_4=\frac{1}{T^2}\iint_{s+t\leqslant T}ds\,dt\int_0^s\mu(s-u)\,dH_1(u)$$
$$\int_s^{s+t}\mu(s+t-v)\,d_vH(v-u)$$
$$\leqslant\Big[\int_0^\infty\mu(x)\,dx\Big]^2\cdot\Big\{\frac{H(T)}{T^2}+\frac{1}{T^2}\int_0^T H(T-x)\,dH(x)\Big\}.$$

In fact, equalities (9) and (11) are obvious, while inequality (10) is obtained on replacing the integration variables $(s, t, x)\to(u, t, x)$, $u=s+t-x$, and substituting a larger domain $\{(u, t, x): 0\leqslant t\leqslant u\leqslant T,\ 0\leqslant x\leqslant T\}$ for the image of the domain of integration. Let us verify inequality (12). After replacing the integration variables $(u, v, s, t)\to(u, w, x, y)$ in accordance with the formulas $s-u=x$, $s+t-v=y$, $v-u=w$, and substituting a larger domain

$$0\leqslant u+w\leqslant T,\ u\geqslant 0,\ w\geqslant 0,\ x\geqslant 0,\ y\geqslant 0$$

for the image of the domain of integration, we obtain

$$I_4\leqslant\int_0^\infty\mu(x)\,dx\cdot\int_0^\infty\mu(y)\,dy\cdot\frac{1}{T^2}\int_0^T dH(w)\int_0^{T-w}dH_1(u)$$
$$=\Big[\int_0^\infty\mu(x)\,dx\Big]^2\cdot\frac{1}{T^2}\int_0^T H_1(T-w)\,dH(w).$$

Now, (12) is derived from the inequality

$$H_1(x)\leqslant A_1(x)+H(x)\leqslant 1+H(x)$$

(see formula (3), § 1, Ch. 4).

6°. It follows from (4)-(8) that

$$E\bar{x}^2(T)\leqslant 2\,[I_1+I_2+I_3+I_4].$$

Since

$$\frac{H_1(T)}{T}\to a,\quad \frac{H(T)}{T}\to a,\quad \frac{1}{T}\int_0^T[1-A_1(x)]\,dx\to 0$$

as $T\to\infty$, to verify inequality (3), it suffices to see that

(13) $$\frac{1}{T}\int_0^T u\,[1-A(u)]\,du\to 0,$$

and

(14) $$\overline{\lim_{T\to\infty}}\frac{1}{T}\int_0^T H(T-x)\,dH(x)\leqslant\frac{a^2}{2}.$$

7°. Statement (13) follows from $T\,[1-A(T)]\to 0$, which, in turn, follows from the relations

$$T\,[1-A(T)]=\int_0^T[1-A(u)]\,du-\int_0^T[A(T)-A(u)]\,du,$$

$$\int_0^T [1-A(u)]\,du \uparrow \int_0^\infty [1-A(u)]\,du = \int_0^\infty du A(u) = a^{-1} < \infty,$$

$$\int_0^T [A(T)-A(u)]\,du \uparrow a^{-1}.$$

8°. We now verify (14). Let $\varepsilon \in (0, 1)$. Put

$$\int_0^T H(T-x)\,dH(x) = I + I_0,\ I_0 = \int_{T-\varepsilon T}^T H(T-x)\,dH(x).$$

Since the function $H(x)$ is non-decreasing,

$$I_0 \leqslant \int_{T-\varepsilon T}^T H(\varepsilon T)\,dH(x) \leqslant H(\varepsilon T)\,H(T). \tag{15}$$

Further, we select a number T_0 such that

$$\left|\frac{H(x)}{x} - a\right| \leqslant \varepsilon \text{ when } \varepsilon x \geqslant T_0.$$

Then, for $\varepsilon T \geqslant T_0$, we have

$$I = \int_0^{T-\varepsilon T} \frac{H(T-x)}{T-x}(T-x)\,dH(x) \leqslant (a+\varepsilon) \int_0^{T-\varepsilon T} (T-x)\,dH(x) \tag{16}$$

$$\leqslant (a+\varepsilon) \int_0^T (T-x)\,dH(x) = (a+\varepsilon) \int_0^T H(x)\,dx;$$

besides,

$$\int_0^T H(x)\,dx = \int_0^{\varepsilon T} H(x)\,dx + \int_{\varepsilon T}^T \frac{H(x)}{x} \cdot x\,dx \leqslant H(\varepsilon T)\cdot \varepsilon T \tag{17}$$

$$+ (a+\varepsilon) \int_{\varepsilon T}^T x\,dx \leqslant H(\varepsilon T)\cdot \varepsilon T + (a+\varepsilon)\,\frac{T^2}{2}.$$

We now have from (15)-(17) for $\varepsilon T \geqslant T_0$:

$$\frac{1}{T^2} \int_0^T H(T-x)\,dH(x) \leqslant \varepsilon \cdot \frac{H(\varepsilon T)}{\varepsilon T} \cdot \frac{H(T)}{T}$$

$$+ (a+\varepsilon)\left\{\varepsilon^2 \frac{H(\varepsilon T)}{\varepsilon T} + \frac{a+\varepsilon}{2}\right\},$$

whence

$$\overline{\lim}\, \frac{1}{T^2} \int_0^T H(T-x)\,dH(x) \leqslant a^2\varepsilon + (a+\varepsilon)\,\varepsilon^2 a + \frac{(a+\varepsilon)^2}{2}$$

as $T \to \infty$, which gives (14) due to the arbitrariness of $\varepsilon \in (0, 1)$.

Chapter 6

Elements of Random Analysis

In order to discuss the continuity of random functions (processes), it is necessary to introduce the concept of convergence of an RV. In probability theory, several types of RV convergence are considered. In this chapter, we define these types of RV convergence, list their basic properties (without proof), consider weak convergence in more detail (in particular, the Prokhorov theorem and the central limit theorem of probability theory, which is related to it) and, finally, certain problems of differentiation and integration.

§ 1. Types of RV Convergence

1°. **Convergence in probability.** A sequence of RV's $\xi_1, \xi_2, \ldots$ converges to an RV ξ in probability (P) if, for each $\varepsilon > 0$,

$$P(|\xi_n - \xi| > \varepsilon) \to 0 \text{ as } n \to \infty.$$

This fact is written in the form

$$\xi_n \xrightarrow{P.} \xi, \text{ or } \operatorname{p\,lim} \xi_n = \xi.$$

In measure theory, this convergence corresponds to convergence in measure.

2°. **Convergence in the mean.** A sequence of RV's $\xi_1, \xi_2, \ldots$ converges to an RV ξ in the mean of order $\lambda > 0$ if

$$E|\xi_n - \xi|^\lambda \to 0 \text{ as } n \to \infty.$$

In the case $\lambda = 2$, this convergence is called convergence in square mean (s.m.), and written in the form

$$\xi_n \xrightarrow{\text{s.m.}} \xi, \text{ or l.i.m. } \xi_n = \xi$$

(l.i.m. = limit in the mean).

3°. **Almost certain convergence.** A sequence of RV's $\xi_1, \xi_2, \ldots$ converges to an RV ξ almost certainly, or with probability unity, if

$$P(\lim \xi_n = \xi) = 1.$$

This fact is written in the form

$$\xi_n \xrightarrow{\text{a.c.}} \xi, \text{ or P lim } \xi_n = \xi.$$

Note that the set

$$\{\lim \xi_n = \xi\} = \{\omega : \lim \xi_n(\omega) = \xi(\omega)\}$$

$$= \bigcap_{m \geqslant 1} \bigcup_{N \geqslant 1} \bigcap_{n \geqslant N} \left\{|\xi_n - \xi| < \frac{1}{m}\right\} \subset \Omega$$

is an event. In measure theory, this convergence corresponds to convergence a.e.

4°. **Convergence in distribution.** A sequence of RV's $\xi_1, \xi_2, \ldots$ converges to an RV ξ_0 in distribution (d.) if

$$F_n(x) \to F_0(x)$$

at all points x of continuity of the (limit) function F_0, where F_k is the DF of the RV ξ_k, $k \geqslant 0$. This fact is often written in the form

$$\xi_n \xrightarrow{\text{d.}} \xi_0.$$

The following relations are true for the types of convergence introduced, viz.,

$$\begin{array}{ccc} \xi_n & \xrightarrow{\text{a.c.}} & \xi \\ & \Downarrow & \\ \xi_n & \xrightarrow{\text{P.}} & \xi \Rightarrow \xi_n \xrightarrow{\text{d.}} \xi \\ & \Uparrow & \\ \xi_n & \xrightarrow{\text{s.m.}} & \xi. \end{array}$$

There are no other relations of this type without additional assumptions.

The following lemmas are often used in applications.

Lemma 1. $\xi_n \xrightarrow{\text{P.}} c = \text{const}$ is equivalent to $\xi_n \xrightarrow{\text{d.}} c$.

Lemma 2. Let f be a continuous function on R^1. Then

(1) $\xi_n \xrightarrow{\text{P.}} \xi$ implies $f(\xi_n) \xrightarrow{\text{P.}} f(\xi)$, and

(2) $\xi_n \xrightarrow{\text{d.}} \xi$ implies $f(\xi_n) \xrightarrow{\text{d.}} f(\xi)$.

This statement is also true in the case of VRV's.

§ 2. Equivalent Definitions of RV Convergence and Basic Properties

1°. We now list the basic statements related to convergence in probability.

Theorem (Ky Fan). For two RV's ξ and η, we put

$$\rho(\xi, \eta) = \inf\{\varepsilon > 0 : P(|\xi - \eta| > \varepsilon) < \varepsilon\}.$$

It turns out that ρ is a metric if we regard RV's coinciding with probability unity as equivalent. Then convergence in probability is equivalent to convergence in metric ρ (see *Traité de statistique théorique et appliqué* by D. Dugué, Masson, Paris, 1958).

Theorem (Slutsky). The space of RV's is complete in the sense of convergence in probability (see the above-mentioned book by D. Dugué).

2°. And now basic statements concerning almost certain convergence.

(1) The space of RV's is complete in the sense of almost certain convergence.

(2) Any sequence of RV's, which is convergent to a certain RV in probability, contains a subsequence convergent to the same RV a.c. (Slutsky theorem).

(3) **(Sufficient Condition for A. C. Convergence).** For a.c. convergence of a sequence of RV's $\xi_1, \xi_2, \ldots$, it is sufficient that there should exist such a summable sequence of positive numbers $\varepsilon_1, \varepsilon_2, \ldots$ that

$$\sum_{n \geqslant 1} P(|\xi_{n+1} - \xi_n| > \varepsilon_n) < \infty.$$

3°. Finally, some basic statements on convergence in distribution.

(1) Convergence in distribution is equivalent to that in each of the following three metrics, viz.,

$$\rho_1(F, G) = \int_{R^1} |F(x) - G(x)|\, \omega(x)\, dx,$$

where $\omega(x) \geqslant 0$ a.e. on R^1, and $\int_{R^1} \omega(x)\, dx < \infty$,

$$\rho_2(F, G) = \int_{R^1} |F(K+x) - G(K+x)|\, dx, \quad K = (0, 1),$$

while $K + x$ is the set obtained by shift through x points from K, whereas $F(K + x) = P(\xi \in K + x)$, ξ being an RV with distribution F, and

$$\rho_3(F, G) = \inf\{\varepsilon > 0 : F(x-\varepsilon) - \varepsilon \leqslant G(x) \leqslant F(x+\varepsilon) + \varepsilon,\ x \in R^1\}.$$

The latter metric is said to be **Prokhorov-Lévy.**

(2) **(Pólya theorem).** If $\{F_n : n \geqslant 0\}$ is a sequence of DF's, the function F_0 is continuous, and $F_n \to F_0$, then $F_n \to F_0$ uniformly.

(3) **(Helly theorem).** Any sequence of DF's contains a subsequence convergent for all values of x (whose limit may not be a DF).

§ 3. Prokhorov Theorem and Its Applications

1°. Let

$(R, \mathbf{B})$ be a Polish (i.e., metric, complete and separable) space, $\mathrm{Cap}(R)$ collection of all probability measures on $(R, \mathbf{B})$, and $C(R)$ the space of real, bounded and continuous functions on R.

We will say that a sequence $\{P_n\} \subset \mathrm{Cap}(R)$ converges **weakly** (w.) to a probability measure $P \in \mathrm{Cap}(R)$ if

$$\int_R f\, dP_n \to \int_R f\, dP \quad \forall f \in C(R).$$

This fact is denoted by

$$P_n \xrightarrow{\text{w.}} P.$$

We will sometimes use the notation

$$(P, f) = \int_R f\, dP.$$

Remark 1. Let $R = R^1 \cdot F_n \xrightarrow{\text{d.}} F_0$ is equivalent to $P_n \xrightarrow{\text{w.}} P_0$, where F_n is a DF, P_n a probability measure corresponding to F_n, and $F_n \xrightarrow{\text{d.}} F_0$ means that $F_n(x) \to F_0(x)$ at the points of continuity of the function F_0.

Remark 2. $(P, f) = (Q, f)$ $\forall$ $f \in C(R)$ implies $P = Q$.

A family $E \subset \text{Cap}(R)$ is said to be **relatively compact** if, for any sequence $\{P_n\} \subset E$, there exists a subsequence $\{P_{n'}\} \subset \{P_n\}$, and $P \in \text{Cap}(R)$, such that $P_{n'} \xrightarrow{\text{w.}} P$ (where P does not necessarily belong to E).

A family $E \subset \text{Cap}(R)$ is said to be **dense** if, for each $\varepsilon > 0$, there exists a compact set K such that $P(K) > 1 - \varepsilon$ for all $P \in E$.

Theorem (Prokhorov). Assume that R is a Polish space. A family $E \subset \text{Cap}(R)$ is relatively compact if and only if it is dense.

Remark 3. A family E is relatively compact if it is dense. In this case (which is important for applications), the metric space may not be necessarily separable and complete.

Remark 4. For P and $Q \in \text{Cap}(R)$, we put

$$\pi(P, Q) = \inf\{\varepsilon > 0 : P(A) \leqslant Q(A^\varepsilon) + \varepsilon,$$
$$Q(A) \leqslant P(A^\varepsilon) + \varepsilon \text{ for all } A \in \mathbf{B}\},$$

where $A^\varepsilon = \{x : \rho(x, A) < \varepsilon\}$. It turns out that π is a metric (it is called the **Prokhorov metric**), and convergence in the metric space $(\text{Cap}(R), \pi)$ coincides with weak convergence.

2°. Convergence of CF's. Let $R = R^s$, $s \geqslant 1$. The CF φ of a probability measure P on R^s is defined by the equality

$$\varphi(t) = \int_{R^s} e^{i(t,x)}\, P(dx),\ t \in R^s \tag{1}$$

(the integral exists, for the function $x \to \exp\{i(t, x)\}$ is continuous and bounded). It is obvious that the function φ is continuous. E-quality (1) defines a mapping $k : P \to \varphi$. It was shown (see § 15, Ch. 2) that k is a bijection.

Theorem. $P_n \xrightarrow{\text{w.}} P_0$ is equivalent to $\varphi_n(t) \to \varphi_0(t)$ $\forall t$, where $\{P_n : n \geqslant 0\} \subset \text{Cap}(R^s)$, and $\varphi_n = k(P_n)$.

Remark 5. If $X = \text{Cap}(R^s)$, $Y = k(X)$, and we consider weak convergence in X, while pointwise in Y (i.e., assume that $\varphi_n \to \varphi_0$ if $\varphi_n(t) \to \varphi_0(t)$ $\forall t$), then this theorem means that the mapping $k : X \to Y$ is a continuous bijection.

Proof. 1°. Let $P_n \xrightarrow{\text{w.}} P_0$. For any $t \in R^s$, we put

$$f(x) = \cos(t, x),\ g(x) = \sin(t, x).$$

Since $f, g \in C(R^s)$, it follows from the definition of weak convergence that

$$\varphi_n(t) = (P_n, f) + i(P_n, g) \to (P_0, f) + i(P_0, g) = \varphi_0(t).$$

2°. Now, let $\varphi_n(t) \to \varphi_0(t)$ $\forall t$. The function φ_0 is continuous, but we shall make use of the continuity of the function φ_0 only at one point $t = 0$; therefore, we derive that the family $E = \{P_1, P_2, \ldots\} \subset \text{Cap}(R^s)$ is dense. It suffices to consider the case $s = 1$ to which the general case is simply reduced. Using the Fubini theorem on changing the order of integration, we have

$$\frac{1}{u}\int_{-u}^{u}[1 - \varphi_n(t)]\, dt = \int_{R^1}\left[\frac{1}{u}\int_{-u}^{u}(1 - e^{itx})\, dt\right] P_n(dx)$$
$$= 2\int_{R^1}\left[1 - \frac{\sin ux}{ux}\right] P_n(dx)$$
$$\geqslant \int_{|x| > \frac{2}{u}} P_n(dx) = P_n\left(x : |x| > \frac{2}{u}\right)$$

for $u > 0$, since

$$2\left(1 - \frac{\sin ux}{ux}\right) \geqslant 2\left(1 - \frac{|\sin ux|}{|ux|}\right) \geqslant 2\left(1 - \frac{1}{|ux|}\right) \geqslant 1$$

when $|ux| \geqslant 2$. The function φ_0 (as the limit of measurable functions bounded by one constant) is a measurable bounded function; therefore, it is integrable. Since, in addition, it is continuous at the point $t = 0$, we obtain that, for $\varepsilon > 0$, there exists $u > 0$, provided

$$\left|\frac{1}{u}\int_{-u}^{u}[1 - \varphi_0(t)]\, dt\right| < \varepsilon.$$

By the Lebesgue theorem,

$$\frac{1}{u}\int_{-u}^{u}[1 - \varphi_n(t)]\, dt \to \frac{1}{u}\int_{-u}^{u}[1 - \varphi_0(t)]\, dt$$

for a fixed number u; therefore,

$$\left|\frac{1}{u}\int_{-u}^{u}[1 - \varphi_n(t)]\, dt\right| \leqslant 2\varepsilon \text{ when } n \geqslant N_\varepsilon, \text{ i.e.,}$$

$$P_n\left(x : |x| > \frac{2}{u}\right) \leqslant 2\varepsilon, \text{ or } P_n\left(x : |x| \leqslant \frac{2}{u}\right) \geqslant 1 - 2\varepsilon$$

when $n \geqslant N_\varepsilon$. It remains to select a number $a \geqslant \frac{2}{u}$ such that

$$P_n(x : |x| \leqslant a) \geqslant 1 - 2\varepsilon\ \forall n.$$

Thus, the family E is dense.

By the Prokhorov theorem, the same family $\{P_n : n \geqslant 1\}$ is relatively compact; hence, any sequence $\{P_{n'}\} \subset \{P_n\}$ contains a subsequence $\{P_{n''}\} \subset \{P_{n'}\}$ and a probability distribution $P \in \mathrm{Cap}\,(R)$ such that $P_{n''} \xrightarrow{\text{w.}} P$. Since $\varphi_{n''} \to \varphi_0 = k\,(P_0)$, and, according to Item 1°, $\varphi_{n''} \to k\,(P)$, it follows from the bijectivity of the mapping k that $P = P_0$; therefore, $P_n \xrightarrow{\text{w.}} P_0$ (which is not complicated to deduce from the following statement, viz., for a sequence of numbers $\{x_n\}$ to be convergent to a number x_0, it is necessary and sufficient that any sequence $\{x_{n'}\} \subset \{x_n\}$ should contain a subsequence $\{x_{n''}\} \subset \{x_{n'}\}$ such that $x_{n''} \to x_0$).

Remark 6. A more general result was proved in Item 2°, viz., if a subsequence of CF's φ_n, $n = 1, 2, \ldots$, corresponding to distributions P_n, $n = 1, 2, \ldots$, converges (pointwise) to a certain function φ which is continuous at zero, then there exists a distribution P such that $P_n \xrightarrow{\text{w.}} P$, and φ is the CF of P.

3° **(Central Limit Theorem of Probability Theory).** Let $\xi_1, \xi_2, \ldots$ be a sequence of i.i.d. RV's. Assume that

$$E\xi_1 = \mu, \quad \operatorname{var} \xi_1 = \sigma^2 < \infty, \quad \sigma > 0,$$

and put

$$S_n = \xi_1 + \ldots + \xi_n,$$

$$\zeta_n = \frac{S_n - ES_n}{\sqrt{\operatorname{var} S_n}} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{\xi_k - \mu}{\sigma}.$$

Theorem.

$$F_n(x) = P(\zeta_n < x) \to \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}}\, du \quad \forall x.$$

Remark 7. By the Pólya theorem (see § 2) and continuity of the function Φ, this convergence is uniform, i.e.,

$$\sup_x |F_n(x) - \Phi(x)| \to 0.$$

Remark 8. If ζ is an RV with normal distribution $N(0, 1)$, then the theorem states that $\zeta_n \xrightarrow{\text{d.}} \zeta$. No other type of convergence occurs in this case. More precisely, there exists no such RV ζ with normal distribution $N(0, 1)$ that

$$\zeta_n \xrightarrow{\text{s.m.}} \zeta, \quad \zeta_n \xrightarrow{\text{a.c.}} \zeta, \quad \text{or} \quad \zeta_n \xrightarrow{\text{P.}} \zeta.$$

Proof. Put $\varphi_n(t) = Ee^{it\zeta_n}$, $t \in R^1$. By the theorem of Item 2°, it suffices to verify that $\varphi_n(t) \to \varphi(t) = e^{-t^2/2}$ (recall that φ is the CF of the standard normal distribution). We can assume that $\mu = 0$,

$\sigma = 1$ (otherwise, we should take $\xi_k' = \frac{\xi_k - \mu}{\sigma}$ instead of ξ_k). Then

$$\zeta_n = \frac{1}{\sqrt{n}} \sum_{1}^{n} \xi_k, \quad \varphi_n(t) = [Ee^{i\frac{t}{\sqrt{n}}\xi_1}]^n = \varphi_1\left(\frac{t}{\sqrt{n}}\right)^n$$

$$= \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{2n}\right)\right]^n \to e^{-\frac{t^2}{2}} \text{ as } n \to \infty.$$

Remark 9. If, in addition to the conditions of the theorem, we assume that

$$\rho = E\,|\xi_1 - \mu|^3 < \infty,$$

then there exists an absolute constant (independent of the type of distribution of the RV ξ_1) $C > 0$ such that

$$|F_n(x) - \Phi(x)| \leqslant C \cdot \frac{1}{\sqrt{n}} \cdot \frac{\rho}{\sigma^3} \quad \forall x.$$

This inequality is called the **Berry-Esseen** inequality. With the efforts of V.M. Zolotarev, the constant C was "made" very small, viz., $C < 1$.

§ 4. Stochastic Continuity

We shall assume that a random function f is defined on a metric space (X, d) with values in a metric space (Y, ρ), and (as always) that the RV's $f(x)$, $x \in X$, are defined on the same probability space (not depending on x).

An RF f is said to be **stochastically continuous** (or **continuous in probability**) at a point $x_0 \in X$ if $\forall \varepsilon > 0$, $\varepsilon_1 > 0$ $\exists \delta > 0$ such that

$$P\{\rho(f(x), f(x_0)) > \varepsilon\} < \varepsilon_1 \text{ when } d(x, x_0) < \delta.$$

The function f is stochastically continuous on a certain set if it is stochastically continuous at each point of this set.

If a function f assumes numerical values (or values in a normed space), then the **stochastic boundedness** of the function means that $\forall \varepsilon > 0$ $\exists N$ such that

$$P\{|f(x)| \leqslant N\} \geqslant 1 - \varepsilon \quad \forall x.$$

In the general case, the latter requirement takes the form

$$P\{\rho(f(x), y_0) \leqslant N\} \geqslant 1 - \varepsilon \quad \forall x$$

for a certain $y_0 \in Y$.

Similarly, **stochastic uniform continuity** is defined, viz., $\forall \varepsilon > 0$ $\varepsilon_1 > 0$ $\exists \delta > 0$ such that

$$P\{\rho(f(x_1), f(x_2)) > \varepsilon\} < \varepsilon_1 \text{ when } d(x_1, x_2) < \delta.$$

Just like for (non-random) analysis, the usual theorems are true, viz., if f is defined on a compactum, and is stochastically continuous, then it is stochastically bounded and uniformly continuous. These

statements are proved in the usual way. An analogue even of the Weierstrass theorem is valid, i.e., a numerical (random) function (stochastically) continuous on [0, 1] is the limit of a (stochastically) uniformly convergent sequence of (random) polynomials (polynomials with random coefficients).

Nevertheless, convergence in probability is not suitable for the further development of random analysis in the spirit of usual (non-random) mathematical analysis, e.g., for deriving statements similar to those of differentiation theory. Thus, if the derivative $\xi'(t)$ of a random function $\xi(t)$ exists, and equals zero on an interval, then it does not follow that the function itself equals a constant a.c., whereas the derivative $\xi'(t)$ is understood as the limit in probability of the RV $\xi_h(t) = \frac{1}{h}[\xi(t+h) - \xi(t)]$ as $h \to 0$ (the Poisson process $\xi(t)$, $t \geqslant 0$, is stochastically continuous, and the indicated derivative exists and equals zero a.c.).

Remark. For constructing random analysis, the use of convergence in square mean proves to be interesting. The usual continuity, differentiation and integration theorems remain valid (certainly, in a new interpretation), while the requirements which should be imposed are only related to the requirements for the two functions

$$\mu(t) = E\,\xi(t) \text{ and } R(s, t) = \operatorname{cov}(\xi(s), \xi(t)).$$

E.g., for a random function $t \to \xi(t)$ to be continuously differentiable (in square mean) on an interval, it is necessary and sufficient that the function $\mu(t)$ should be continuously differentiable, while $R(s, t)$ should have continuous mixed derivative.

§ 5. Stochastic Integrals of Non-Random Functions

The goal of this section is to define integrals of the form

$$\int_T f(t)\, d\xi(t),$$

where f is a function on a finite or infinite interval $T \subset R^1$, and $\{\xi(t) : t \in T\}$ a process with uncorrelated increments

$$E\,|\,\xi(t)\,|^2 < \infty, \quad E\xi(t) = 0 \ \forall t.$$

Remark. Such a process (with uncorrelated increments) has been chosen, first, because such integrals may be encountered in the theory of stationary (in the wide sense) processes, and, second, if the process $\xi(t)$, $t \in T$, were differentiable (in the sense of a certain type of convergence), then the problem would be reduced to the definition of an integral of the form

$$\int_T [f(t)\,\xi'(t)]\, dt$$

(i.e., an integral of a random function with respect to a non-random measure). The latter problem is more simple (e.g., the latter integral may be defined as an integral along each trajectory of the process); however, the process with uncorrelated increments is not differentiable. Therefore, such a reduction is impossible. In fact, if there existed a derivative $\eta = \xi'(t)$ in the sense of convergence in square mean, i.e.,

$$E\left|\frac{1}{h}[\xi(t+h) - \xi(t)] - \eta\right|^2 \to 0 \text{ as } h \to 0,$$

then, in particular,

$$\delta = E\left|\frac{1}{h_2}[\xi(t+h_2) - \xi(t)] - \frac{1}{h_1}[\xi(t+h_1) - \xi(t)]\right|^2 \to 0$$

when $0 < h_1 < h_2 \to 0$. However, assuming for simplicity $T = [a, b]$, $\xi(a) = 0$, and $F(t) = \operatorname{var}\xi(t)$, $t \in T$, we derive from the formula

$$E\,|\,\xi(t) - \xi(s)\,|^2 = F(t) - F(s) \text{ when } t > s, \tag{1}$$

which is obtained from the uncorrelatedness of increments $\xi(s)$ and $\xi(t) - \xi(s)$, that $(0 < h_1 < h_2)$

$$\delta = E\left|\frac{1}{h_2}[\xi(t+h_2) - \xi(t+h_1)] + \left(\frac{1}{h_2} - \frac{1}{h_1}\right)[\xi(t+h_1) - \xi(t)]\right|^2$$

$$= \frac{1}{h_2^2}[F(t+h_2) - F(t+h_1)] + \left(\frac{1}{h_2} - \frac{1}{h_1}\right)^2 [F(t+h_1) - F(t)],$$

the latter expression being not convergent to zero as $h_2, h_1 \to 0$ (e.g., for a strictly monotonic function F and when $h_2 = 2h_1 \downarrow 0$, the second addend on the right-hand side of the latter formula may converge only to $+\infty$).

Let $T = [a, b]$ or $T = R^1$.

Definition. A complex-valued random process $\{\xi(t) : t \in T\}$ is called a **process with uncorrelated increments** if

(1) $E\,|\,\xi(t)\,|^2 < \infty$, $E\,\xi(t) = 0\ \forall t$, and

(2) for any $t_1 \leqslant t_2 \leqslant t_3 \leqslant t_4$ from T,

$$E[\xi(t_2) - \xi(t_1)] \cdot \overline{[\xi(t_4) - \xi(t_3)]} = 0.$$

Put

$$F(t) = E\,|\xi(t)\,|^2 = \operatorname{var}\xi(t), \quad t \in T.$$

The function F is called the **structure function** of the process. In the case $T = R^1$, we will usually assume that

$$F(-\infty) = 0, \text{ and } F(+\infty) < \infty. \tag{2}$$

In particular, this implies formula (1) if we use the uncorrelatedness of the two addends on the right-hand side when $u < s < t$, and $\lim_{u \to -\infty} E\,|\,\xi(u)\,|^2 = 0$ in the equality

$$\xi(t) - \xi(u) = [\xi(s) - \xi(u)] + [\xi(t) - \xi(s)];$$

therefore, the structural function F is always non-decreasing (and bounded).

(1) and (2) imply the existence of such two RV's $\xi(-\infty)$ and $\xi(+\infty)$ with finite second moments and zero means that

$$\xi(t) \xrightarrow{\text{s.m.}} \xi(-\infty) = 0 \text{ as } t \to -\infty,$$

$$\xi(t) \xrightarrow{\text{s.m.}} \xi(+\infty) \text{ as } t \to +\infty.$$

In fact, $\xi(t) \xrightarrow{\text{s.m.}} 0$ as $t \to -\infty$ is equivalent to $F(-\infty) = 0$. It follows from (1) and (2) that $E\,|\,\xi(t) - \xi(s)\,|^2 \to 0$ as $s, t \to +\infty$. It remains to use the completeness of the space of (complex) RV's with finite second moments and zero means.

Theorem 1. Let $\{\xi(t) : t \in T\}$ be a (complex-valued) random process with uncorrelated increments.

(1) If $T = [a, b]$, then, for any (complex-valued) continuous function f on T, there exists a (complex) RV ξ_f such that $E|\xi_f|^2 < \infty$, $E\,\xi_f = 0$, and

$$E\Big|\sum_{i=1}^{n} f(u_i)\,[\xi(t_i) - \xi(t_{i-1})] - \xi_f\Big|^2 \to 0 \tag{3}$$

when $a = t_0 < t_1 \ldots < t_n = b$, $t_{i-1} \leqslant u_i \leqslant t_i$, $\max\limits_i (t_i - t_{i-1}) \to 0$.

(2) If $T = (-\infty, \infty)$, the structural function F is bounded, and $F(-\infty) = 0$, then, for any continuous, bounded function f on T, there exists an RV ξ_f such that $E\,|\,\xi_f\,|^2 < \infty$, $E\xi_f = 0$, and (3) is fulfilled when $-\infty = t_0 < t_1 < \ldots < t_n = +\infty$, $t_{i-1} \leqslant u_i < t_i$, $\max\limits_{1<i<n} (t_i - t_{i-1}) \to 0$, $t_1 \to -\infty$, $t_{n-1} \to +\infty$.

This theorem permits us to give a definition of the stochastic integral of a (non-random) continuous, bounded function. It is natural that the RV ξ_f is unique with probability unity. By definition, we put

$$\xi_f = \int_T f(t)\,d\xi(t) = \int_T f\,d\xi.$$

Note that if $T = [a, b]$, then we can always assume $\xi(a) = 0$ (otherwise, we should pass to the process $\eta(t) = \xi(t) - \xi(a)\ \forall t$).

Corollary.

$$\operatorname{var} \int_T f\,d\xi = \int_T |f(t)|^2\,dF(t). \tag{4}$$

Moreover, if, in addition, g is a continuous, bounded function on T, and

$$\xi_g = \int_T g\,d\xi, \qquad \text{then}$$

$$\operatorname{cov}(\xi_f, \xi_g) = E\xi_f \cdot \overline{\xi}_g = \int_T f(t)\,\overline{g(t)}\,dF(t). \tag{5}$$

It suffices to verify formula (4) to which (5) is a corollary. Put

$$S_n(f) = \sum_1^n f(u_i)\,[\xi(t_i) - \xi(t_{i-1})].$$

Then we derive from the uncorrelatedness of the increments of the process $\xi(t)$, $t \in T$, that

$$\operatorname{var} S_n(f) = ES_n(f) \cdot \overline{S_n(f)} = \sum_{i=1}^{n} f(u_i) \cdot \overline{f(u_i)}\,E\,|\xi(t_i) - \xi(t_{i-1})|^2$$

$$= \sum_1^n |f(u_i)|^2\,[F(t_i) - F(t_{i-1})].$$

It remains to make use of Theorem 1 and the fact that the latter integral sum converges to integral (4).

A structural function F generates on $(T, \mathbf{B})$ a finite σ-additive measure m by the equality $m(\alpha, \beta] = F(\beta + 0) - F(\alpha + 0)$, $\alpha < \beta$, where $\mathbf{B}$ is the σ-algebra of Borel subsets of the set T. Let $L^2 = L^2(T, \mathbf{B}, m)$ be the space of measurable (complex-valued) square-integrable (with respect to measure m) functions on $(T, \mathbf{B})$, C the space of continuous functions bounded on T, and H the space of (complex) RV's with finite second moments and zero means. It is clear that $C \subset L^2$.

Theorem 2. There exists a unique mapping $J : L^2 \to H$ with the following properties, viz.,

(1) J is linear,

(2) J is isometric, i.e.,

$$\operatorname{var} J(f) = E\,|\,J(f)\,|^2 = \int_T |f|^2 dF;$$

moreover,

$$\operatorname{cov}(J(f), F(g)) = EJ(f)\,\overline{J(g)} = \int_T f\bar{g}\,dF$$

for f and g from L^2,

(3) $\|f_n - f\|_{L^2} \to 0$ implies $J(f_n) \xrightarrow{\text{s.m.}} J(f)$ for f, f_n from L^2, and

(4) the mapping J coincides on $C \subset L^2$ with the integral defined by Theorem 1, i.e.,

$$J(f) = \int_T f\,d\xi \tag{6}$$

for $f \in C$.

This theorem makes it possible to give the definition of the stochastic integral of a (non-random) function from L^2 by means of equality (6). Note that, by isometricity, formulas (4)-(5) are now already valid for all f and g from L^2.

Proof of Theorem 1. Consider the proof of the theorem only for the case $T = [a, b]$, since, by the finiteness of the measure m generated

by the function F, the case $T = R^1$ is reduced to the first (or considered similarly). Let $\{A_1, \ldots, A_n\}$ be a finite decomposition of the set $T = [a, b]$ into disjoint intervals (open, half-open and closed), and $u_1, \ldots, u_n$ points from $A_1, \ldots, A_n$, respectively. Put $\alpha = \{u_i, A_i\}$ and

$$S_\alpha = \sum_i f(u_i)\,\xi(A_i),$$

where $\xi(A_i)$ is the increment of the process on the interval A_i (i.e., the difference of the values of the process at the ends of this interval). We call the object α a (finite) decomposition. If $\beta = \{v_j, B_j\}$ is another decomposition, then

$$S_\alpha - S_\beta = \sum_{i,j} f(u_i)\,\xi(A_iB_j) - \sum_{i,j} f(v_j)\,\xi(A_iB_j)$$
$$= \sum_{i,j} [f(u_i) - f(v_j)]\,\xi(A_iB_j).$$

Note that A_iB_j are also intervals, though some of them are empty. We derive from (1) and uncorrelatedness of the increments that

$$E\,|S_\alpha - S_\beta|^2 = \sum_{i,j} |f(u_i) - f(v_j)|^2\,F(A_iB_j).$$

If $A_iB_j = \varnothing$, then, certainly, $F(A_iB_j) = 0$; therefore, if all the intervals A_i, B_j are of length not greater than $\delta > 0$, then, taking into account that $A_iB_j \neq \varnothing$ implies $|u_i - v_j| \leqslant 2\delta$, we have

$$E\,|S_\alpha - S_\beta|^2 \leqslant \varepsilon^2 \sum_{i,j} F(A_iB_j) = \varepsilon^2 F([a, b]),$$

where

$$\varepsilon = \sup_{|u-v|\leqslant 2\delta} |f(u) - f(v)|.$$

Thus, if

$$|\alpha| = \max_i |A_i| \to 0, \quad |\beta| = \max_j |B_j| \to 0$$

(where $|A|$ is the length of an interval A), then it follows from the uniform continuity of the function f that

$$E\,|S_\alpha - S_\beta|^2 \to 0.$$

It remains to consider the space of RV's with finite second moments (i.e., the space of numerical functions square-integrable with respect to measure P on (Ω, S) if (Ω, S, P) is the initial probability space), and make use of the completeness of this space, from which it follows that there exists such an RV ξ_f with finite second moment that

$$E\,|S_\alpha - \xi_f|^2 \to 0 \text{ as } |\alpha| \to 0.$$

Since $ES_\alpha = 0$, we derive from the relation

$$|E\xi_f|^2 = |E(S_\alpha - \xi_f)|^2 \leqslant E\,|S_\alpha - \xi_f|^2$$

that $E\xi_f = 0$.

Proof of Theorem 2 is given in the next section (see Example 3).

§ 6. Orthogonal Measure and Integral with Values in Hilbert Space

A. Let H be a Hilbert space, and E a certain set. In this section, we deal with measures on subsets of the set E, with values in H, and with integrals with respect to such measures, also with values in H. First, we give a series of definitions.

Denote the algebra and σ-algebra of subsets of the set E by $\mathcal{F}_0$ and $\mathcal{F}$, $\mathcal{F}_0 \subset \mathcal{F}$, and the scalar product of two elements a and b from H by $[a, b]$.

D1. A mapping $\mu: \mathcal{F}_0 \to H$ is called a **measure** if

$$\mu(A + B) = \mu(A) + \mu(B)$$

for any two disjoint sets A and B from $\mathcal{F}_0$.

D2. A measure $\mu: \mathcal{F}_0 \to H$ is said to be **finite** if

$$\|\mu(A)\| < \infty \text{ for all } A \in \mathcal{F}_0.$$

D3. A measure μ of $\mathcal{F}_0$ is said to be **σ-additive (countably additive)** if, for such disjoint sets $A_1, A_2, \ldots$ from $\mathcal{F}_0$ that $A = \sum_{k\geqslant 1} A_k \in \mathcal{F}_0$, the equality holds:

$$\mu(A) = \sum_{k=1} \mu(A_k),$$

i.e.,

$$\Big\|\mu(A) - \sum_{k=1}^{n} \mu(A_k)\Big\| \to 0 \text{ as } n \to \infty.$$

D4. A measure μ on $\mathcal{F}_0$ is said to be **orthogonal** if

$$[\mu(A), \mu(B)] = 0$$

for any two disjoint sets A and B from $\mathcal{F}_0$.

Let a measure $\mu: \mathcal{F}_0 \to H$ be orthogonal, σ-additive and finite. The natural question is asked, viz., does there exist a unique extension of the measure μ from $\mathcal{F}_0$ to the minimal σ-algebra $\mathcal{F} \supset \mathcal{F}_0$, preserving the orthogonality, σ-additivity and finiteness properties? The answer to this question is positive: we shall now give the construction of such an extension, and simultaneously the construction of integrals with respect to this measure.

B. Put

$$m(A) = \|\mu(A)\|^2 \quad \text{for} \quad A \in \mathcal{F}_0.$$

Then $m: \mathcal{F}_0 \to [0, \infty)$ is the usual finite, σ-additive measure on the algebra $\mathcal{F}_0$, and there exists a unique extension of this measure from $\mathcal{F}_0$ to $\mathcal{F}$, which we shall also denote by the symbol m, and which is also a finite, σ-additive measure (we may even assume that m-zero sets are added to $\mathcal{F}$, so that the space $(E, \mathcal{F}, m)$ with measure is complete).

Denote the space of measurable functions square-integrable on $(E, \mathcal{F})$ with respect to measure m by $L^2 = L^2(E, \mathcal{F}, m)$. The functions may be either complex-valued (but then the Hilbert space H must be complex) or real-valued (and then the space H may be real or complex). For f and g from L^2, the scalar product is

$$(f, g) = \int_E f\bar{g}\, dm, \quad \| f \| = (f, f)^{1/2},$$

where $\bar{g}(x) = \overline{g(x)}$ is the complex conjugate to the number $g(x)$.

C. Denote by L_0^2 the set of functions from L^2 of the form

$$f = \sum c_k 1_{A_k}, \quad A_k \in \mathcal{F}_0$$

with a finite number of addends, where c_k is a number, and 1_A the indicator of the set A. Note that the linear subspace $L_0^2 \subset L^2$ is dense in L^2, i.e., for any $f \in L^2$, there exists a sequence $\{f_n\} \subset L_0^2$ such that $\| f_n - f \| \to 0$ as $n \to \infty$.

We now define a mapping $J : L_0^2 \to H$ by putting

$$J(f) = \sum c_k \mu(A_k) \quad \text{for} \quad f = \sum c_k 1_{A_k} \in L_0^2.$$

Note that this mapping J is isometric, i.e., $\| J(f) \| = \| f \| \; \forall f \in L_0^2$, or, which is tantamount,

$$[J(f), J(g)] = (f, g) \; \forall f, \quad g \in L_0^2.$$

D. Extend the mapping J from L_0^2 to L^2, preserving linearity and isometry. For $f \in L^2$, choose $\{f_n\} \subset L_0^2$ so that $\| f_n - f \| \to 0$. Since J is isometric on L_0^2,

$$\| J(f_n) - J(f_m) \| = \| f_n - f_m \| \to 0 \text{ as } n, m \to \infty.$$

Since the space H is complete, and the sequence $\{J(f_n)\} \subset H$ fundamental, there exists such an element in H, which we denote by $J(f)$, that

$$\| J(f_n) - J(f) \| \to 0.$$

It is obvious that $J(f)$ does not depend on the choice of a sequence $\{f_n\}$ with the indicated property. Put

(1) $$J(f) = \int_E f\, d\mu.$$

It is clear that the mapping $J: L^2 \to H$ is linear and isometric, i.e.,

(2) $$[J(f), J(g)] = (f, g),$$

(3) $$J(af + bg) = aJ(f) + bJ(g)$$

for all f, g from L^2 and all numbers a and b. Certainly, it is not necessary that $J(L^2) = H$.

E. We now make use of the definition of integral (1) for the construction of the extension of measure μ from $\mathcal{F}_0$ to $\mathcal{F}$. For any $A \in \mathcal{F}$,

the function $1_A \in L^2$. Put

$$\mu^*(A) = J(1_A) \quad \text{for} \quad A \in \mathcal{F}.$$

We show that the mapping $\mu^* : \mathcal{F} \to H$ so defined is an orthogonal, σ-additive and finite measure on $\mathcal{F}$, which coincides with μ on $\mathcal{F}_0$. In fact,

(1) $$[\mu^*(A), \mu^*(B)] = [J(1_A), J(1_B)] = (1_A, 1_B) = 0$$

for any two disjoint sets A and B from $\mathcal{F}$,

(2) for pairwise disjoint sets $A_1, A_2, \ldots$ from $\mathcal{F}$, we have

$$A = \sum_{k \geqslant 1} A_k \in \mathcal{F},$$

$$\| \mu^*(A) - \sum_{k=1}^{n} \mu^*(A_k) \| = \| J(1_A) - \sum_{k=1}^{n} J(1_{A_k}) \|$$

$$= \| J(1_A - \sum_{1}^{n} 1_{A_k}) \| = \| J(1_{\sum_{k>n} A_k}) \|$$

$$= \| 1_{\sum_{k>n} A_k} \| = \{ m(\sum_{k>n} A_k) \}^{1/2} \to 0 \text{ as } n \to \infty,$$

(3) $$\| \mu^*(A) \| = \| J(1_A) \| = \| 1_A \| = \{m(A)\}^{1/2} < \infty,$$

and

(4) for $A \in \mathcal{F}_0$, the function $1_A \in L_0^2$; therefore,

$$\mu^*(A) = J(1_A) = \mu(A).$$

For the extension μ^* of the measure μ from $\mathcal{F}_0$ to $\mathcal{F}$, we will employ the same notation as before, i.e., $\mu^* = \mu$.

Example 1. Let $f \in L^2(E, \mathcal{F}, m)$. Define a mapping $\lambda : \mathcal{F} \to H$ by putting

$$\lambda(A) = \int_E 1_A(x) f(x)\, \mu(dx), \quad A \in \mathcal{F}.$$

Then

P1. λ is an orthogonal, σ-additive and finite measure on $(E, \mathcal{F})$ with values in H,

P2. If $g \in L^2(E, \mathcal{F}, m_f)$, which is equivalent to $fg \in F^2(E, \mathcal{F}, m)$, where the measure $m_f : \mathcal{F} \to [0, \infty)$ is defined by the equality

$$m_f(A) = \| \lambda(A) \|^2 = \int_A | f(x) |^2 m(dx),$$

then

(4) $$\int_E g(x) \lambda(dx) = \int_E g(x) f(x) \mu(dx).$$

We now verify P1.

(1) For two disjoint sets A and B from $\mathcal{F}$, we have

$$[\lambda(A),\ \lambda(B)] = (1_A f,\ 1_B f) = \int_E 1_A \cdot 1_B \cdot |f|^2\, dm = 0.$$

(2) If $A = \sum_{k\geqslant 1} A_k$, where $\{A_k\} \subset \mathcal{F}$, then

$$\left\| \lambda(A) - \sum_1^n \lambda(A_k) \right\|^2 = \left\| \int_E 1_{A \setminus \sum_1^n A_k} f\, d\mu \right\|^2$$

$$= \left\| \int_E 1_{\sum_{k>n} A_k} f\, d\mu \right\|^2 = \int_E 1_{\sum_{k>n} A_k} |f|^2\, dm$$

$$= \int_{\sum_{k>n} A_k} |f|^2\, dm \to 0 \text{ as } n \to \infty.$$

(3) $$\| \lambda(A) \|^2 = \int_E 1_A |f|^2\, dm \leqslant \int_E |f|^2\, dm < \infty.$$

To verify P2, we notice that $g \in L^2(E, \mathcal{F}, m_f)$ is equivalent to

$$\int_E |g(x)|^2\, m_f(dx) = \int_E |g(x)|^2\, |f(x)|^2\, m(dx),$$

which, in turn, is equivalent to $fg \in L^2(E, \mathcal{F}, m)$. It now remains to verify formula (4).

(1) For functions g assuming a finite number of values, i.e., having the form

(5) $$g = \sum c_k 1_{A_k},\ \{A_k\} \in \mathcal{F},\ \bigcup A_k = E$$

with a finite number of addends, we put $I(g) = \int_E g(x)\, \lambda(dx)$ for $g \in L^2(E, \mathcal{F}, m_f)$. Then we derive from the definition of I that $I(1_A) = \lambda(A)$; hence,

$$I(g) = \sum c_k I(1_{A_k}) = \sum c_k \lambda(A_k) = \sum c_k J(1_{A_k} f)$$
$$= J\left(\sum c_k 1_{A_k} f\right) = J(gf).$$

(2) Note first of all that convergence $g_n \to g$ in the norm of the space $L^2(E, \mathcal{F}, m_f)$ is equivalent to convergence $g_n f \to gf$ in the norm of the space $L^2(E, \mathcal{F}, m)$, since

$$\int_E |g_n - g|^2\, dm_f = \int_E |g_n - g|^2\, |f|^2\, dm.$$

Now, if $\{g_n\}$ is a sequence of functions of form (5), and $g_n \to g$ in the norm of the space $L^2(E, \mathcal{F}, m_f)$, then it follows from the relations

$$I(g_n) \to I(g)$$
$$\| \qquad\quad \|$$
$$J(g_n f) \to J(gf)$$

(in the norm of the space H) that $I(g) = J(gf)$, i.e., formula (4). It remains to take into account the fact that the set of functions of form (5) is everywhere dense in $L^2(E, \mathcal{F}, m_f)$.

Example 2. Let $(E, \mathcal{F}) = (R^1, \mathbf{B}^1)$. The usual finite measure m on the straight line is in one-to-one correspondence with a non-decreasing, left-continuous and bounded function F such that

$$m[a, b) = F(b) - F(a).$$

A similar statement is also valid in the case of an orthogonal, σ-additive and finite measure μ.

D5. We call a function, $\xi : R^1 \to H$ a **function, or process with orthogonal increments** if, for any $t_1 \leqslant t_2 \leqslant t_3 \leqslant t_4$,

$$[\xi(t_2) - \xi(t_1),\quad \xi(t_4) - \xi(t_3)] = 0.$$

Let μ be an orthogonal, σ-additive and finite measure with the structural measure

$$m(A) = \| \mu(A) \|^2, \quad A \in \mathcal{F} = \mathbf{B}^1$$

and structural function

$$F(t) = m(-\infty, t), \quad m[a, b) = F(b) - F(a).$$

Putting

$$\xi(t) = \mu(-\infty, t), \quad t \in R^1,$$

we derive that

(1) the function $\xi : R^1 \to H$ is a function with orthogonal increments,

(2) $$\mu[a, b) = \xi(b) - \xi(a), \quad F(t) = \| \xi(t) \|^2,$$

and

(3) F is a bounded (positive semi-definite, non-decreasing and left-continuous) function, with $F(-\infty) = 0$.

Conversely, let $\xi : R^1 \to H$ be a function with orthogonal increments and bounded structural function $F(t) = \| \xi(t) \|^2$ such that $F(-\infty) = 0$. The function F is non-decreasing, since

(6) $$\| \xi(t) - \xi(s) \|^2 = F(t) - F(s) \text{ when } t > s.$$

In fact, since

$$\lim_{u \to -\infty} \| \xi(u) \|^2 = F(-\infty) = 0,$$

we have

$$\lim_{u \to -\infty} \| \xi(t) - \xi(u) \| = \| \xi(t) \|\ \forall t,$$

whereas we obtain from the orthogonality of increments that

$$\| \xi(t) - \xi(u) \|^2 = \| [\xi(s) - \xi(u)] + [\xi(t) - \xi(s)] \|^2$$
$$= \| \xi(s) - \xi(u) \|^2 + \| \xi(t) - \xi(s) \|^2$$

when $u < s < t$, which implies (6) as $u \to -\infty$.

We now associate the function ξ with a function ξ_+: $R^1 \to H$ by putting

$$\xi_+(t) = \xi(t + 0)$$

(in the sense of convergence in the norm of the space H). This definition is correct, since

$$\| \xi(t + h_2) - \xi(t + h_1) \|^2 = F(t + h_2) - F(t + h_1) \to 0$$

when $0 < h_1 \leqslant h_2 \to 0$; therefore, due to the completeness of H, there exists such an element $\xi_+(t)$ in H that

$$\| \xi(t + h) - \xi_+(t) \| \to 0 \text{ when } h \downarrow 0.$$

ξ_+ is a function with orthogonal increments, which follows from

$$\xi(t + h) \to \xi_+(t) \text{ when } h \downarrow 0 \ \forall\, t,$$
$$[\xi(t_2 + h) - \xi(t_1 + h), \quad \xi(t_4 + h) - \xi(t_3 + h)] = 0$$

when $t_1 \leqslant t_2 \leqslant t_3 \leqslant t_4$.

Note that the structural function $F_+(t) = \| \xi_+(t) \|^2$ of the process ξ_+ is right-continuous, and $F_+(t) = F(t + 0)$. The process ξ_- with structural function F_- is defined similarly.

Since there exist

$$\lim_{t \to \infty} \xi_\pm(t) = \xi(+\infty) \text{ and } \lim_{t \to -\infty} \xi_\pm(t) = \xi(-\infty) = 0$$

(which follows from (6), the boundedness of F, equality $F(-\infty) = 0$ and completeness of the space H), as $(E, \mathcal{F})$, we can take $E = R^1$ with accessory point $-\infty$ and/or $+\infty$ and the corresponding σ-algebra $\mathcal{F}$ containing $\mathbf{B}^1$ and the accessory point (points). Let $E = [-\infty, \infty)$, and $\mathcal{F}_0$ be the algebra of sets of the form $A = \sum [a_k, b_k)$ with a finite number of addends, $-\infty \leqslant a_k \leqslant b_k \leqslant +\infty$. Putting

$$m[a, b) = F_-(b) - F_-(a), \quad m(A) = \sum m[a_k, b_k),$$
$$\mu[a, b) = \xi_-(b) - \xi_-(a), \quad \mu(A) = \sum \mu[a_k, b_k),$$

we obtain that

(1) the mapping $m: \mathcal{F}_0 \to [0, \infty)$ is finite, σ-additive measure uniquely extendable from $\mathcal{F}_0$ to $\mathcal{F} \supseteq \mathbf{B}^1$, and preserving the finiteness and σ-additivity properties,

(2) the mapping μ: $\mathcal{F}_0 \to H$ is a measure which is

(a) orthogonal, since the increments of the function ξ_- are orthogonal,

(b) finite and even such that $\| \mu(A) \|^2 = \sum \| \mu[a_k, b_k) \|^2 = \sum \| \xi_-(b_k) - \xi_-(a_k) \|^2 = \sum [F_-(b_k) - F_-(a_k)] = \sum m[a_k, b_k) = m(A)$,

(c) σ-additive, since if sets $A_1, A_2, \ldots$ from $\mathcal{F}_0$ are disjoint, and

$$A = \sum_{k \geqslant 1} A_k \in \mathcal{F}_0,$$

then we have

$$\mu(A) - \sum_1^n \mu(A_k) \|^2 = \| \mu(A \setminus \sum_1^n A_k) \|^2 = \| \mu \sum_{k > n} A_k \|^2$$
$$= m(\sum_{k > n} A_k) \to 0 \text{ as } n \to \infty.$$

Such a measure μ is uniquely extendable from $\mathcal{F}_0$ to $\mathcal{F}$, and preserves the finiteness, σ-additivity and orthogonality properties.

Similarly to the notation

$$\int_{R^1} f\, dF = \int_{R^1} f\, dm,$$

we define the integral $\int_{R^1} f\, d\xi = \int_{-\infty}^{\infty} f(t)\, d\xi(t)$ by the equality

$$\int_{R^1} f\, d\xi = \int_{R^1} f\, d\mu \text{ for } f \in L^2(R^1, \mathbf{B}^1, m).$$

Remark. If, however, the function F is unbounded, but $F(-\infty) = 0$, then the measure m is σ-finite (i.e., representable as the union of a countable number of disjoint sets $E_1, E_2, \ldots$, whereas $m(E_k) < \infty \ \forall\, k$), and the usual constructions of measure and integral are suitable.

Example 3. Let H be the space of (complex or real) RV's on a probability space (Ω, S, P) with finite second moments, zero means and

$$[\xi, \eta] = E\xi\bar{\eta} = \operatorname{cov}(\xi, \eta) \text{ for } \xi, \eta \in H$$

as the scalar product. If $T = [a, b]$ or $T = (-\infty, \infty)$, then the function $\xi: T \to H$ with orthogonal increments is, in other words, a random process with uncorrelated increments, in which case Theorem 2 of the previous section follows from the results of Example 2. It only remains to prove statement (4).

If F is (left-) continuous at a point t, then $\xi(t) = \xi_-(t)$ a.c., since

$$\| \xi(t) - \xi_-(t) \| \leqslant \| \xi(t) - \xi(t - h) \| + \| \xi(t - h) - \xi_-(t) \|,$$

and

$$\| \xi(t - h) - \xi_-(t) \| \to 0,$$
$$\| \xi(t) - \xi(t - h) \|^2 = F(t) - F(t - h) \to 0$$

when $h \downarrow 0$.

The process $\xi(t)$, $t \in T$, can be replaced in Theorem 1 by the process $\xi_-(t)$, $t \in T$, while the value of the integral $\int_T f(t)\, d\xi(t)$ defined by this theorem a.c. will not change. In fact, the points u_i are selected arbitrarily in the interval from t_{i-1} to t_i. Taking them from the set of points of continuity of the function F, we obtain that the integral sums for the processes ξ and ξ_- involved in Theorem 1 coincide a.c.; therefore, their limits coincide a.c.

In this connection, we can assume that $\xi(t) = \xi_-(t)\ \forall t$. Denoting the sets

$$[t_0, t_1), \quad [t_1, t_2), \quad \ldots, \quad [t_{n-1}, t_n)$$

by $A_1, A_2, \ldots, A_n$, respectively, we have

$$J(1_{A_k}) = \mu(A_k) = \xi(t_k) - \xi(t_{k-1})\ \forall\, k = 1, \ldots, n;$$

therefore,

$$S_n = \sum_{k=1}^{n} f(u_k)\,[\xi(t_k) - \xi(t_{k-1})] = J\Big(\sum_1^n f(u_k)\, 1_{A_k}\Big) = J(f_n),$$

where $f_n = \sum_1^n f(u_k)\, 1_{A_k}$. Since $\| f_n - f \| \to 0$, $J(f_n) \xrightarrow{\text{s.m.}} J(f)$. Then, by Theorem 1,

$$J(f) = \int_T f(t)\, d\xi(t).$$

The case $T = [a, b]$ is investigated similarly.

Example 4. Let T be a finite or infinite interval, and $\xi_0\colon T \to H$ a process with orthogonal increments and structure function $F_0(t) = \| \xi_0(t) \|^2$, $t \in T$. If f on T is complex-valued and square-integrable with respect to structural measure determined uniquely by the function F_0, then the process

$$T \ni t \to \xi(t) = \int_{t_0}^{t} f(\tau)\, d\xi_0(\tau) \in H$$

has orthogonal increments, where $\int_{t_0}^{t} = -\int_{t}^{t_0}$ when $t < t_0$, which follows from the following, viz.,

(1) this statement is valid for functions of the form $f = 1_A$, $A = [a, b)$, and their finite linear combinations $f = \Sigma f_k 1_{A_k}$ with disjoint A_k,

and

(2) if $\xi_n\colon T \to H$, $n \geqslant 1$, is a sequence of functions with orthogonal increments, and $\xi_n(t) \to \xi(t) \in H$ (in the norm of the space H) for each $t \in T$, then the limit function $\xi\colon T \to H$ has orthogonal increments.

The structural function F of a process $\{\xi(t), t \in T\}$ is defined by the equality

$$F(t) = \| \xi(t) \|^2 = \int_{t_0}^{t} |f(\tau)|^2\, dF_0(\tau),$$

which is first proved by the verification for the indicators $f = 1_A$, $A = [a, b)$, and then for their linear combinations $f = \Sigma f_k 1_{A_k}$ with disjoint A_k.

§ 7. Differentiation and Integration of Mathematical Expectation of Processes

The goal of the section is to indicate conditions under which the following formulas are valid, viz.,

$$\frac{d}{dt} E\xi(t) = E\frac{d}{dt}\xi(t), \qquad \int_a^b E\xi(t)\, dt = E\int_a^b \xi(t)\, dt$$

for a process $\{\xi(t) : t \in R^1\}$. Differentiation and integration are performed for each trajectory separately (i.e., in the sense of almost certain convergence). In other words, if $\{\xi(t, \omega) : t \in R^1\}$ is a trajectory of the process under consideration, which corresponds to an elementary event ω, then

$$\Big(\frac{d}{dt}\xi(t)\Big)(\omega) = \frac{\partial}{\partial t}\xi(t, \omega), \quad \text{and} \quad \Big(\int_a^b \xi(t)\, dt\Big)(\omega) = \int_a^b \xi(t, \omega)\, dt.$$

Theorem 1. Consider a real random process $\{\xi(t) : t \in R^1\}$. Let

(1) the RV's $\xi(t)$ be integrable ($E\,|\,\xi(t)\,| < \infty\ \forall\, t$),

(2) $\xi(t, \omega)$ differentiable with respect to t for all ω, and

(3) the derivative $\xi'(t, \omega) = \frac{\partial}{\partial t}\xi(t, \omega)$ satisfy the condition

$$|\,\xi'(t, \omega)\,| \leqslant \xi(\omega),$$

where ξ is an integrable RV. Then the function $E\xi(t)$ is differentiable with respect to t, and

$$\frac{d}{dt} E\xi(t) = E\xi'(t).$$

Theorem 2. Let a process $\{\xi(t) : a \leqslant t \leqslant b\}$ satisfy the conditions

(1) $|\,\xi(t)\,| \leqslant \xi$, where ξ is an integrable RV,

(2) $\xi(t, \omega)$ is integrable with respect to t in the sense of Riemann for all ω,

and

(3) $E\xi(t)$ is integrable with respect to t in the sense of Riemann.

Then

$$\int_a^b E\xi(t)\,dt = E\int_a^b \xi(t)\,dt.$$

Proof of Theorem 1. Put

$$\eta(h) = \frac{1}{h}[\xi(t+h) - \xi(t)], \quad h > 0$$

for a fixed t.

Then

(1) the RV's $\eta(h)$ are integrable $(E\,|\,\eta(h)\,| < \infty\ \forall\, h > 0)$,

(2) $|\,\eta(h)\,| \leqslant \xi$,

since

$$\eta(h, w) = \xi'(t + \theta_\omega \cdot h, w), \quad 0 \leqslant \theta_\omega \leqslant 1,$$
$$|\,\eta(h, w)| \leqslant |\sup_{0\leqslant\theta\leqslant 1} |\,\xi'(t + \theta h, w)| \leqslant \xi(w),$$

and

(3) $\eta(h) \to \xi'(t)$ a.c. when $h \downarrow 0$.

Therefore, by the Lebesgue theorem,

(1) $\xi'(t)$ is an integrable RV,

and

(2) $E\eta(h) \to E\xi'(t)$ when $h \downarrow 0$,

but

$$E\eta(h) = \frac{1}{h}[E\xi(t+h) - E\xi(t)] \to \frac{d}{dt}E\xi(t)$$

(since $\lim_{h\downarrow 0} E\eta(h)$ exists, and the case $h < 0$ is considered similarly).

Proof of Theorem 2. Let

$$S_n = \frac{1}{n}\sum_{k=1}^{n} \xi(a + kh), \quad h = \frac{b-a}{n}.$$

We have

(1) S_n is an integrable RV,

(2) $|\,S_n\,| \leqslant \xi$, where ξ is an integrable RV,

and

(3) $S_n \to S = \int_a^b \xi(t)\,dt$.

Therefore, by the Lebesgue theorem,

(1) S is an integrable RV,

and

(2) $ES_n \to ES = E\int_a^b \xi(t)\,dt$.

However,

$$ES_n = \frac{1}{n}\sum_{1}^{n} E\xi(a + kh) \to \int_a^b E\xi(t)\,dt.$$

Chapter 7

Gaussian Processes

§ 1. Definition of Gaussian Process

A random process $\{\xi(t) : t \in T\}$ is said to be **Gaussian** if all its finite-dimensional distributions are Gaussian, i.e., normal. We call a distribution normal also in the case where the covariance matrix A is positive semi-definite.

Remark. Such a distribution exists and equals the weak limit of normal distributions with positive definite covariance matrix $A + \varepsilon I$ when $\varepsilon \downarrow 0$, in which case the corresponding sequence of characteristic functions converges to a function which is continuous at zero (see Remark 6 from § 3, Ch. 6).

Put

$$\mu(t) = E\xi(t), \quad R(s, t) = \operatorname{cov}(\xi(s), \quad \xi(t)).$$

The function μ is called a mean value (or mathematical expectation, or shift function), and function R the **covariance function** of the process. However, R is more often said to be the **correlation function** apparently, to underline that the function R reflects the dependence (correlation) of the values of the process at various points.

Note that, for any finite set $\{t_1, \ldots, t_n\} \subset T$, the matrix $\{R(t_i, t_j,)\}$ is positive semi-definite. We now formulate the converse statement.

Theorem. Consider two arbitrary numerical functions μ and R on T and $T \times T$, respectively, and assume that, for any finite set $\{t_1, \ldots, t_n\} \subset T$, the matrix $\{R(t_i, t_j)\}$ is positive semi-definite. Then there exists a Gaussian process $\{\xi(t) : t \in T\}$ with shift function μ and covariance function R.

This proposition immediately follows from the Kolmogorov theorem (see Introduction to Part II), since the family of probability measures P_M, $M = \{t_1, \ldots, t_n\} \subset T$, satisfies the consistency condition, P_M denoting a measure corresponding to the normal distribution with mean vector

$$\mu_M = \{\mu(t_1,), \ldots, \mu(t_n)\}$$

and covariance matrix

$$A_M = \{R(t_i, t_j)\}.$$

§ 2. Wiener Process

A. A homogeneous Gaussian process $\{\xi(t) : t \geqslant 0\}$ with independent increments is said to be **Wiener**, or **Brownian motion process**, in which case $T = [0, \infty)$, while a process is said to be **homogeneous** if the distribution of the RV $\xi(s + t) - \xi(s)$ depends only on $t \geqslant 0$ (does not depend on $s \geqslant 0$), and it is said to be a **process**

with independent increments if, for any numbers $0 = t_0 < t_1 < \ldots < t_n$, the RV's

$$\xi(t_0),\ \xi(t_1) - \xi(t_0),\ \ldots,\ \xi(t_n) - \xi(t_{n-1})$$

are independent. We shall assume that $\xi(0) = 0$ (otherwise, we can take the process $\eta(t) = \xi(t) - \xi(0),\ t \geqslant 0$).

B. We show that if $\xi(t),\ t \geqslant 0$, is a Wiener process with measurable function $t \to E\xi(t)$, then we have

$$E\xi(t) = at,\ \operatorname{var}\xi(t) = \sigma^2 \cdot t,$$

$$R(s, t) = \operatorname{cov}(\xi(s), \xi(t)) = \sigma^2 \cdot \min(s, t)$$

for certain numbers $a,\ \sigma > 0$, and all numbers $s \geqslant 0,\ t \geqslant 0$. The numbers a and σ are called the **shift** and **diffusion coefficient**, respectively, of the Wiener process.

Putting

$$w_t = w(t) = \frac{\xi(t) - at}{\sigma},$$

we get

$$Ew_t = 0,\ \operatorname{var} w_t = t,$$

$$R_0(s, t) = \operatorname{cov}(w_s, w_t) = \min(s, t).$$

Such a process (i.e., a Wiener process with $a = 0$ and $\sigma = 1$) is said to be **standard** Wiener.

Denote the CF of an RV $\xi(t)$ by φ_t. Since

$$\xi(s + t) = \xi(s) + [\xi(s + t) - \xi(s)],$$

it follows from the homogeneity and independence of increments that

$$\varphi_{s+t} = \varphi_s \cdot \varphi_t$$

(pointwise); whence, for each $\lambda \in R^1$, we have

$$\varphi_t(\lambda) = [\varphi_1(\lambda)]^t,$$

which gives

$$\varphi_t(\lambda) = \exp\left\{ia\lambda t - \frac{\sigma^2\lambda^2}{2}t\right\}$$

for certain numbers a and $\sigma > 0$. Thus, the RV $\xi(t)$ has the normal distribution with mean at and variance $\sigma^2 t$. We show that the covariance function is of the form

$$R(s, t) = \operatorname{cov}(\xi(s), \xi(t)) = \sigma^2 \cdot \min(s, t)$$

(note that the $R_0(s, t) = \min(s, t)$ is positive semi-definite, i.e, for any numbers $0 \leqslant t_1 \leqslant \ldots \leqslant t_n$, the matrix $\{R_0(t_i, t_j)\}$ is positive semi-definite). In fact, if $s < t$, then it follows from the homogeneity and independence of increments that

$$\begin{aligned} R(s, t) &= \operatorname{cov}(\xi(s), \xi(s) + [\xi(t) - \xi(s)]) \\ &= \operatorname{cov}(\xi(s), \xi(s)) + \operatorname{cov}(\xi(s), \xi(t) - \xi(s)) \\ &= \operatorname{cov}(\xi(s), \xi(s)) = \operatorname{var}\xi(s) = \sigma^2 s = \sigma^2 \cdot \min(s, t). \end{aligned}$$

C. There exists a modification of a Wiener process $\xi(t),\ t \geqslant 0$ (i.e., a process $\xi'(t),\ t \geqslant 0$, stochastically equivalent to $\xi(t),\ t \geqslant 0$, meaning that $P\{\xi'(t) = \xi(t)\} = 1\ \forall\, t$) whose almost all trajectories are continuous, which follows from the following theorem given without proof.

Kolmogorov Theorem. Let $\{\xi_t : t \in [a, b]\}$ be a random process defined on a complete probability space, and there exist positive constants α, β and C such that

$$E\,|\,\xi_t - \xi_s\,|^\beta \leqslant C\,|\,t - s\,|^{1+\alpha}.$$

Then there exists such a modification of the process that almost all of its trajectories are continuous with respect to t.

In our case, if $\xi_t = w_t,\ t \geqslant 0$, is a standard Wiener process, then $w_t - w_s$ has the normal distribution $N(0,\ |\,t - s\,|)$, whence we obtain

$$E\,|\,w_t - w_s\,|^4 = 3\,|\,t - s\,|^2,$$

and the Kolmogorov theorem is valid for $\alpha = 1$, $\beta = 4$ and $C = 3$.

When we speak of a standard Wiener process, we often mean this very modification.

E. Though we may assume that almost all trajectories of a Wiener process are continuous, the Wiener process is not differentiable at any point (even!) in the sense of convergence in probability. In fact, for any $t > 0$, convergence in probability of an RV $\eta_h = \frac{1}{h}(w_{t+h} - w_t)$ to another when $h \downarrow 0$ would imply convergence in distribution, but the RV η_h has the normal distribution $N(0, h^{-1})$; therefore,

$$P(\eta_h < x) \to F(x) = 0\ \forall\, x.$$

§ 3. Non-Homogeneous Wiener Process

A Gaussian process $\{\xi(t) : t \geqslant 0\}$ with independent increments is called a **non-homogeneous Wiener process**. We shall again assume that $\xi(0) = 0$ (otherwise, we could pass to the process $\eta(t) = \xi(t) - \xi(0),\ t \geqslant 0$).

Theorem. 1°. If $\{\xi(t) : t \geqslant 0\}$ is a non-homogeneous Wiener process, then the covariance function is of the form

$$R(s, t) = \operatorname{cov}(\xi(s), \xi(t)) = F(\min(s, t)), \tag{1}$$

where F is a positive semi-definite, non-decreasing function on $[0, \infty)$. Note that $F(t) = \operatorname{var}\xi(t),\ t \geqslant 0$.

2°. If F is a positive semi-definite, non-decreasing function on $[0, \infty)$, then the function R of form (1) on $[0, \infty) \times [0, \infty)$ is positive semi-definite (recall that this means the positive semi-definiteness of the matrix $\{R(t_i, t_j)\}$ for any finite set $\{t_1, \ldots, t_n\} \subset T = [0, \infty)$). This implies that a Gaussian process with covariance function of form (1) does exist.

3°. A Gaussian process with covariance function of form (1) is a process with independent increments, i.e., a non-homogeneous Wiener process.

Thus, a non-homogeneous Wiener process $\{\xi(t): t \geqslant 0\}$ exists, and is fully determined by the two functions

$$\mu(t) = E\xi(t) \text{ and } F(t) = \operatorname{var} \xi(t), \ t \geqslant 0,$$

where, according to the assumption that $\xi(0) = 0$, we have $\mu(0) = 0$, $F(0) = 0$. Besides, the function F is non-decreasing.

Proof. 1°. Due to the independence of increments, we have

$$\begin{aligned} R(s, t) &= \operatorname{cov}(\xi(s), \ \xi(s) + [\xi(t) - \xi(s)]) \\ &= \operatorname{cov}(\xi(s), \ \xi(s)) + \operatorname{cov}(\xi(s), \ \xi(t) - \xi(s)) \\ &= \operatorname{cov}(\xi(s), \ \xi(s)) = \operatorname{var} \xi(s) \\ &= F(s) = F(\min(s, t)) \geqslant 0 \end{aligned}$$

when $s < t$. From the independence of increments again and equality

$$\xi(t) = \xi(s) + [\xi(t) - \xi(s)],$$

we derive for $s < t$ that

$$\operatorname{var} \xi(t) = \operatorname{var} \xi(s) + \operatorname{var} [\xi(t) - \xi(s)] \geqslant \operatorname{var} \xi(s),$$

i.e., $F(t) \geqslant F(s)$ when $t \geqslant s$.

2°. Let $0 \leqslant t_1 \leqslant t_2 \leqslant \ldots \leqslant t_n$. It should be shown that the matrix

$$R = \{R_{ij}\}, \text{ where } R_{ij} = R(t_i, t_j) = F(\min(t_i, t_j)) = F_{\min(i,j)},$$

is positive semi-definite. We have from the data that $F_1 \leqslant F_2 \leqslant \ldots \leqslant F_n$. Put

$$\Delta_1 = F_1, \ \Delta_2 = F_2 - F_1, \ \ldots, \ \Delta_n = F_n - F_{n-1},$$

$$C = \{c_{ij}\} = \left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ \hline 1 & 1 & 1 & 1 \end{array}\right], \qquad \Lambda = \{\delta_{ij}\Delta_i\}.$$

Then $R = C\Lambda C'$. Indeed,

$$c_{ij} = \begin{cases} 1 & \text{when } i \geqslant j \\ 0 & \text{when } i < j \end{cases}$$

$$\begin{aligned} R_{ij} &= \sum_{\alpha, \beta} c_{i\alpha}\delta_{\alpha\beta}\Delta_\alpha c_{j\beta} = \sum_{\alpha} c_{i\alpha}\Delta_\alpha c_{j\alpha} \\ &= \sum_{\substack{\alpha \leqslant i \\ \alpha \leqslant j}} \Delta_\alpha = \sum_{\alpha < \min(i, j)} \Delta_\alpha = F_{\min(i, j)}. \end{aligned}$$

Since the diagonal matrix Λ consists of non-negative elements, the matrix R is positive semi-definite.

3°. We show that the increments of the process are pairwise uncorrelated on disjoint intervals. Let $s < t < u < v$. Then

$$\begin{aligned} &\operatorname{cov}(\xi(t) - \xi(s), \ \xi(v) - \xi(u)) = R(t, v) + R(s, u) \\ &- R(s, v) - R(t, u) = F(t) + F(s) - F(s) - F(t) = 0. \end{aligned}$$

It remains to make use of the following statement, viz., for normally distributed RV's, uncorrelatedness is equivalent to independence, and pairwise independence is equivalent to mutual independence.

Remark 1. The theorem remains valid if we select an arbitrary totally ordered set instead of $T = [0, \infty)$ (e.g., any subset of the straight line with the natural ordering).

Remark 2. Assume that $\mu(t) = E\xi(t) = 0 \ \forall \, t \in T = [0, \infty)$. Since

$$\begin{aligned} \operatorname{var} [\xi(t) - \xi(s)] &= R(t, t) + R(s, s) - R(t, s) - R(s, t) \\ &= F(t) + F(s) - F(s) - F(s) = F(t) - F(s) \end{aligned}$$

when $s < t$, we have

$$E \, |\xi(t) - \xi(s)|^2 = F(t) - F(s) \text{ when } t \geqslant s.$$

E.g., if a non-decreasing function F is right-continuous, then the non-homogeneous Wiener process (without shift) is right-continuous in the sense of convergence in square mean.

Remark 3. Let $\{w(t) : t \geqslant 0\}$ be a standard Wiener process, a numerical function f positive semi-definite on $[0, \infty)$, and the Lebesgue integral $\int_0^\infty f(x)\,dx < \infty$. Put

$$\xi(t) = \int_0^t \sqrt{f(\tau)}\, dw(\tau), \tag{2}$$

$$F(t) = \int_0^t f(x)\,dx, \quad t \geqslant 0.$$

Then $\{\xi(t) : t \geqslant 0\}$ is a non-homogeneous Wiener process with structural function $F(t) = \operatorname{var} \xi(t)$, $t \geqslant 0$, and zero shift ($E\xi(t) = 0 \ \forall \, t$), which follows from the results of Example 4, § 6, Ch. 6, and the fact that linear transformation (2) changes a Gaussian process into Gaussian.

Thus, if the structural function of a non-homogeneous Wiener process possesses density, then such a process is linearly expressed in terms of a standard Wiener process.

Chapter 8

Stationary Processes

The correlation theory of stationary (in the wide sense) random processes is a special part of the theory of Hilbert spaces. After the definition, we give some examples of such processes and a linear representation of a stationary process in terms of a process with orthogonal increments; we also consider linear transformations of the processes (differentiation and integration).

§ 1. Definition of Stationary Process (in Wide Sense). Examples

A process $\{\xi_t : t \in T\}$ is said to be **stationary** (in the wide sense) if

$$E \mid \xi_t \mid^2 < \infty, \quad E\xi_t = E\xi_0,$$

$$\operatorname{cov}(\xi_{s+t}, \ \xi_s) = \operatorname{cov}(\xi_t, \ \xi_0) \ \forall s,t \in T,$$

where $T = \mathbf{Z} = \{0, \pm 1, \pm 2, \ldots\}$ or $T = R^1 = (-\infty, \infty)$, so that $s + t \in T$ for all $s, t \in T$. The cases $T = \{0, 1, 2, \ldots\}$ and $T = [0, \infty)$ are reduced to the previous by extending the process. We assume that the process $\{\xi_t\}$ is complex-valued. It will be seen that this assumption does not complicate, but, on the contrary, simplifies the theory of such processes. By definition,

$$\operatorname{cov}(\xi, \ \eta) = E(\xi - E\eta)\,\overline{(\eta - E\eta)}.$$

Put

$$m = E\xi_0, \quad R(t) = \operatorname{cov}(\xi_t, \ \xi_0), \quad \rho(t) = \frac{R(t)}{R(0)} = \operatorname{cor}(\xi_t, \ \xi_0).$$

The functions R and ρ are said to be **covariance** and **correlation**, respectively. We assume that $m = 0$.

Example 1. Let

$$\xi_t = \zeta_0 \cos \lambda t + \zeta_1 \sin \lambda t, \quad t \in T = R^1 \text{ or } \mathbf{Z},$$

the RV's ζ_0 and ζ_1 uncorrelated, and $E\zeta_i = 0$, $E \mid \zeta_i \mid^2 = 1$ for $i = 0, 1$, in which case

$$\begin{aligned}\operatorname{cov}(\xi_{s+t} \ \xi_s) &= \cos \lambda (s + t) \cdot \cos \lambda s + \sin \lambda (s + t) \cdot \sin \lambda s \\ &= \cos \lambda t = R(t);\end{aligned}$$

therefore, the process $\{\xi_t\}$ is stationary (in the wide sense).

Example 2. Consider a sequence of RV's

$$\xi_t = c_t \cdot \zeta, \ t \in T = \mathbf{Z},$$

where $\{c_t\}$ are complex numbers, $c_0 = 1$, $E\zeta = 0$, and $E \mid \zeta \mid^2 = \sigma^2 > 0$. The condition

$$R(t) = \operatorname{cov}(\xi_{s+t}, \ \xi_s) = \operatorname{cov}(\xi_t, \ \xi_0)$$

is fulfilled for all s and t from T if and only if there exists a real number λ such that $c_t = e^{i\lambda t}$ for all t. Thus, the process

$$\xi_t = e^{i\lambda t} \cdot \zeta, \ t \in \mathbf{Z}$$

is stationary, and $R(t) = \sigma^2 e^{i\lambda t}$.

Example 3. Now let

$$\xi_t = \sum_1^n \zeta_k e^{i\lambda_k t}, \quad t \in T = R^1 \text{ or } \mathbf{Z},$$

where $\{\lambda_k\}$ are real numbers, $\{\zeta_k\}$ uncorrelated RV's with $E\zeta_k = 0$, and $E \mid \zeta_k \mid^2 = \sigma_k^2 \ \forall k$, in which case

$$\operatorname{cov}(\xi_{s+t}, \ \xi_s) = \sum e^{i\lambda_k (s+t)} \overline{e^{i\lambda_k s}} \cdot E\zeta_k \overline{\zeta}_k = \sum_{k=1}^n \sigma_k^2 \cdot e^{i\lambda_k t}$$

does not depend on s; therefore, the process $\{\xi_t\}$ is stationary with covariance function

$$R(t) = \sum_1^n \sigma_k^2 \cdot e^{i\lambda_k t}, \ t \in T.$$

Example 4. Put

$$\xi_t = e^{i(\xi t + \eta)}, \ t \in R^1,$$

where the RV's ξ and η are independent, ξ has DF F, and η is uniformly distributed on $(0, 2\pi)$. Then

$$E\xi_t = \int_{-\infty}^{\infty} \int_0^{2\pi} e^{i(ut+v)} \, dF(u) \cdot \frac{dv}{2\pi} = 0,$$

$$\operatorname{cov}(\xi_{s+t}, \ \xi_s) = Ee^{i[\xi(s+t)+\eta]} \cdot \overline{e^{i(\xi s+\eta)}} = Ee^{it\xi} = \int_{-\infty}^{\infty} e^{itx} \, dF(x).$$

Thus, the process $\{\xi_t\}$ is stationary (in the wide sense) with covariance function

$$R(t) = \int_{-\infty}^{\infty} e^{itx} \, dF(x), \ t \in R^1.$$

§ 2. Representation of Stationary Process in Terms of Process with Uncorrelated Increments

Denote by H the space of complex RV's with finite second moments and zero means. The scalar product

$$(\xi, \eta) = E\xi\bar{\eta}, \quad \xi, \eta \in H$$

transforms this space into Hilbert (space of classes of equivalent RV's coinciding a.c.). Meanwhile, the uncorrelatedness of RV's with zero means turns into orthogonality, and a stationary (in the wide sense) process $\{\xi_t: t \in T\}$ with zero mean is a mapping $\xi: T \to H$ such that

$$(*) \qquad (\xi_{s+t}, \ \xi_s) = (\xi_t, \ \xi_0) \ \forall \ s, t \in T$$

(for convenience of notation, we sometimes write ξ_t instead of $\xi(t)$).

In this connection, we regard H in the sequel as an arbitrary complete Hilbert space, while investigating functions $\xi: T \to H$ which satisfy condition (*), and are said to be **stationary**. Recall that

$$T = \mathbf{Z} = \{0, \pm 1, \pm 2, \ldots\}$$

or

$$T = R^1 = (-\infty, \infty).$$

Put $R(t) = (\xi_t, \xi_0)$. We still call R a covariance function. It is positive semi-definite, i.e., for any complex numbers $c_1, \ldots, c_n$ and $t_1, \ldots, t_n \in T$,

$$\sum_{i,j} c_i \bar{c}_j R(t_i - t_j) \geqslant 0,$$

which is due to the fact that the latter sum equals

$$\| \sum_{i=1}^{n} c_i \xi_{t_i} \|^2.$$

We shall need the following two statements.

Theorem 1 (Herglotz). For a function R on $\mathbf{Z}$ to be positive semi-definite, it is necessary and sufficient that it should be representable in the form

$$(1) \qquad R(n) = \int_{-\pi}^{\pi} e^{inx} m(dx), \ n \in \mathbf{Z},$$

where m is a finite (uniquely determined) measure on $(E, \mathbf{B})$, $E = [-\pi, \pi)$, and $\mathbf{B}$ the class of Borel sets contained in E.

Theorem 2 (Bochner, Khinchine). For a continuous function R on $R^1 = (-\infty, \infty)$ to be positive semi-definite, it is necessary and sufficient that it should be representable in the form

$$(2) \qquad R(t) = \int_{-\infty}^{\infty} e^{itx} m(dx), \ t \in R^1,$$

where m is a finite (uniquely determined) measure on $(E, \mathbf{B})$, $E = R^1$, and $\mathbf{B}$ the class of Borel sets.

Note that formulas (1)-(2) are written in the form

$$(3) \qquad R(t) = \int_{E} e^{itx} m(dx), \ t \in T,$$

where $E = [-\pi, \pi)$ or $(-\infty, \infty)$ according as $T = \mathbf{Z}$ or $(-\infty, \infty)$.

Basic theorem. Let a covariance function R be representable in form (3), where the measure m is finite. Then there exist an orthogonal, σ-additive and finite measure $\mu: \mathbf{B} \to H$ and such corresponding function $\xi: E \to H$ with orthogonal increments [$\zeta(t) = \mu(-\infty, t)$ if $T = R^1$, and $\zeta(t) = \mu[-\pi, t)$ if $T = [-\pi, \pi) \ \forall \ t \in T$) that

$$(4) \qquad \xi_t = \int_E e^{itx} \mu(dx) = \int_E e^{itx} d\zeta(x), \ t \in T,$$

with $m(A) = \| \mu(A) \|^2$. Representation (4) is called the **spectral representation** of the stationary function $\{\xi_t\}$, and the function $F(t) = \| \zeta(t) \|^2$ a **spectral function**. The measure m is uniquely determined by the spectral function via the equality $m[a, b) = F(b) - F(a)$.

Proof. A. Let $L^2 = L^2(E, \mathbf{B}, m)$, $\{e_t: t \in T\} \subset L^2$, where $e_t(x) = e^{itx} \ \forall \ x \in E$, and L_0^2 a linear manifold generated by the functions e_t, $t \in T$. The closure of L_0^2 (under the norm of the space L^2) coincides with L^2 (which is proved in the last item).

In addition, let H_0 be a linear manifold generated by elements $\xi_t \in H$, $t \in T$, and H_1 the closure of H_0 (under the norm of the space H).

We establish a one-to-one correspondence "$\leftrightarrow$" between the elements of L^2 and H_1 by putting $e_t \leftrightarrow \xi_t$, then extend it by linearity ($\sum c_t e_t \leftrightarrow \sum c_t \xi_t$, where the number of addends is finite) between the elements of L_0^2, H_0, and by continuity between the elements of L^2, H_1. Such a definition of correspondence is correct, since we verify before each stage of its extension that it preserves scalar products (and, therefore, norms). In fact,

$$(e_s, e_t) = \int_E e_s(x) \overline{e_t(x)} \, m(dx) = \int_E e^{i(s-t)x} m(dx)$$

$$= R(s-t) = (\xi_s, \xi_t).$$

Similarly,

$$(\sum c_s e_s, \ \sum c_t e_t) = (\sum c_s \xi_s, \ \sum c_t \xi_t),$$

where the number of addends in the sums is finite. Thus, the one-to-one correspondence "$\leftrightarrow$" is linear and isometric.

B. For any Borel set $A \in \mathbf{B}$, the indicator $1_A \in L^2$; therefore, it is associated with a certain element $\mu(A)$ from H_1: $1_A \leftrightarrow \mu(A)$. Since $\| 1_A \|^2 = m(A)$, we obtain, due to isometricity, that $\| \mu(A) \|^2 = m(A)$. For two disjoint sets A an B from $\mathbf{B}$,

$$1_{A+B} \leftrightarrow \mu(A+B),$$
$$1_{A+B} = 1_A + 1_B \leftrightarrow \mu(A) + \mu(B),$$

whence $\mu(A+B) = \mu(A) + \mu(B)$, i.e., $\mu : \mathbf{B} \to H$ is a measure which is

(1) finite, since $\| \mu(A) \| = m(A)^{1/2} < \infty$ for all $A \in \mathbf{B}$,

(2) orthogonal, since, by isometricity, $(\mu(A), \mu'(B)) = (1_A, 1_B) = 0$ for any two disjoint A and B from $\mathbf{B}$, and

(3) σ-additive, since

$$\| \mu(A) - \sum_1^n \mu(A_k) \|^2 = \| \mu(\sum_{k>n} A_k) \|^2 = m(\sum_{k>n} A_k) \to 0$$

as $n \to \infty$, where $A = \sum_{k \geqslant 1} A_k$.

C. Thus, the integral

$$J(f) = \int_E f(x)\, \mu(dx)$$

is defined for $f \in L^2$.

If, for a mapping $S : L^2 \to H_1$ defined by the correspondence $f \leftrightarrow S(f)$, we show that

(5) $$S(f) = J(f) \quad \forall f \in L^2,$$

then, in particular, formula (4) will follow from $S(e_t) = \xi_t$. But formula (5) follows from the fact that the mappings S and J are linear, continuous, and coincide on the indicators (i.e., $S(1_A) = J(1_A)$ for $A \in \mathbf{B}$), whereas finite linear combinations of the indicators compose an everywhere dense set in L^2.

D. It remains to prove that the closure of L_0^2 coincides with L^2. In fact, in the case $E = [-\pi, \pi)$, and, therefore, $T = \mathbf{Z}$, by extending the functions from L^2 at the point π by their values at the point $-\pi$, we obtain that the set of bounded functions on $[-\pi, \pi]$ is everywhere dense in L^2, with the set of continuous (bounded) functions being everywhere dense in the former, and containing, in turn, the set of twice continuously differentiable functions each of which is expanded in a uniformly convergent Fourier series. As to the case $E = (-\infty, \infty)$, and, therefore, $T = (-\infty, \infty)$, the set of bounded functions is everywhere dense in L^2, while in the former set, the set of bounded functions with compact support is everywhere dense (they vanish outside a certain finite interval), the set of continuous functions with compact support being everywhere dense

in the latter set, and containing the set of everywhere dense twice continuously differentiable functions. However, if f is such a function, and vanishes outside an interval $[-l, l]$, then f expands in a uniformly convergent Fourier series on the interval $[-l, l]$.

Example. If $T = (-\infty, \infty)$, and $\{\xi_t : t \in T\}$ is a stationary Gaussian process with zero mean, then the process $\{\zeta(t) : t \in E = T\}$ involved in spectral representation (4) is a non-homogeneous Wiener (complex-valued) process with structural function $F(t) = \operatorname{var} \zeta(t)$, $t \in T$, which follows from the fact that H_0 consists of normally distributed, complex-valued RV's (i.e., complex RV's whose imaginary and real parts have the normal distribution). But then H_1 will also consist of normally distributed complex RV's (since if RV's $\xi_1, \xi_2, \ldots$ have the normal distribution, and $\xi_n \xrightarrow{\text{s.m.}} \xi$, the RV ξ also has the normal distribution). Thus, $\{\zeta(t) : -\infty < t < \infty\}$ is a Gaussian process with uncorrelated increments, i.e., a non-homogeneous Wiener process.

§ 3. Differentiation of Stationary Processes

In the present section, we consider the (linear) differentiation operation transforming stationary processes (in the wide sense) into stationary. Our goal is to clarify conditions under which this operation is applicable, and also how the spectral representation and covariance function alter.

Consider a stationary process $\{\xi_t : t \in T = (-\infty, \infty)\}$ with values in a complex Hilbert space H. Assume that the covariance function admits the representation

$$R(t) = (\xi_{s+t}, \xi_s) = \int_{-\infty}^{\infty} e^{itx} m(dx), \quad t \in T,$$

where the (spectral) measure m on $(R^1, \mathbf{B}^1)$ is finite. Let

$$\xi_t = \int_{-\infty}^{\infty} e^{itx} \mu(dx), \quad t \in T$$

be the spectral representation.

Theorem. Assume that

(1) $$\int_{-\infty}^{\infty} |x|^{2r} m(dx) < \infty.$$

Let P be a polynomial of degree r (with complex coefficients).

Then

(1) for any integer k, $1 \leqslant k \leqslant r$, the random function $\xi : t \to \xi_t$ is differentiable k times (in the sense of convergence in the norm of the space H),

(2) the mapping

$$T \ni t \to \eta_t = (P(D)\,\xi)(t) \in H$$

is a stationary process, where D is the differentiation operation (in the sense of convergence in the norm of the space H),

(3) the process $\{\eta_t : t \in T\}$ has spectral representation of the form

$$\eta_t = \int_{-\infty}^{\infty} e^{itx} \mu_\eta(dx) \tag{2}$$

and covariance function of the form

$$R_\eta(t) = \int_{-\infty}^{\infty} e^{itx} m_\eta(dx), \tag{3}$$

where the measures μ_η and m_η are determined in terms of μ and m by the equalities

$$\mu_\eta(A) = \int_A P(ix)\,\mu(dx) = \int_{-\infty}^{\infty} 1_A(x) \cdot P(ix)\,\mu(dx),$$

$$m_\eta(A) = \|\mu_\eta(A)\|^2 = \int_A |P(ix)|^2\, m(dx), \quad A \in \mathbf{B},$$

(4) the covariance function R is differentiable (continuously) 2^r times, whereas

$$R_\eta = P(D)\,\overline{P}(-D)\,R. \tag{4}$$

Proof. A. We first show that there exists the derivative

$$(D\xi)(t) = \xi_t' = \lim_{\substack{h \to 0 \\ h \neq 0}} \frac{1}{h}[\xi_{t+h} - \xi_t],$$

where

$$\xi_t' = \int_{-\infty}^{\infty} e^{itx} \cdot ix\mu(dx), \quad -\infty < t < \infty,$$

for which we consider the expression

$$\frac{1}{h}[\xi_{t+h} - \xi_t] = \int_{-\infty}^{\infty} e^{itx} \frac{e^{ihx} - 1}{h}\,\mu(dx), \quad h \neq 0,$$

and make use of the inequality

$$\left|\frac{e^{ihx} - 1}{h}\right| \leqslant |x|.$$

Since $f_h \in L^2 = L^2(R^1, \mathbf{B}^1, m)$, while

$$f_h(x) = e^{itx} \frac{e^{ihx} - 1}{h},$$

and $f_h \to f$ in the norm of the space L^2, where $f(x) = e^{itx} \cdot ix$, we have $J(f_h) \to J(f)$ in the norm of the space H (see § 6, Ch. 6), i.e., $\lim \frac{1}{h}[\xi_{t+h} - \xi_t]$ exists as $h \to 0$, $h \neq 0$, and equals

$$J(f) = \int_{-\infty}^{\infty} e^{itx} ix\mu(dx).$$

We now verify by induction that, for each integer k, $1 \leqslant k \leqslant r$, the function $t \to \xi_t$ is k times differentiable, while

$$(D^k\xi)(t) = \xi_t^{(k)} = \int_{-\infty}^{\infty} e^{itx} (ix)^k \mu(dx). \tag{5}$$

For $k = 1$, this formula has already been proved. Let it be valid for a certain $k < r$. Consider the expression

$$\frac{1}{h}[\xi_{t+h}^{(k)} - \xi_t^{(k)}] = \int_{-\infty}^{\infty} e^{itx} \frac{e^{ihx} - 1}{h} (ix)^k \mu(dx), \quad h \neq 0,$$

and make use of the fact that

$$|f_h(x)| \leqslant |x|^{k+1}$$

for the function

$$f_h(x) = e^{itx} \frac{e^{ihx} - 1}{h} (ix)^k, \quad -\infty < x < \infty.$$

Since $f_h \in L^2$, and $f_h \to f$ in the norm of the space L^2, where $f(x) = e^{itx} \times (ix)^{k+1}$, we again have $J(f_h) \to J(f)$ as $h \to 0$, $h \neq 0$, in the norm of the space H, i.e., there exists $\lim \frac{1}{h}[\xi_{t+h}^{(k)} - \xi_t^{(k)}]$ equal to

$$J(f) = \int_{-\infty}^{\infty} e^{itx} (ix)^{k+1} \mu(dx).$$

B. The process $\eta_t = (P(D)\,\xi)(t)$, $-\infty < t < \infty$ is now defined, and, according to (5),

$$\eta_t = \int_{-\infty}^{\infty} e^{itx} P(ix)\,\mu(dx).$$

Putting

$$f(x) = P(ix), \quad g(x) = e^{itx},$$

we make use of Example 1, § 6, Ch. 6. Equality (2) holds due to the result of this example, and the measure m_η is orthogonal, σ-additive and finite. Formula (3) is obtained from the isometricity of the mapping

$$L^2(R^1, \mathbf{B}^1, m_\eta) \ni f \to J_\eta(f) \in H,$$

since

$$R_\eta(t) = (\eta_{s+t}, \eta_s) = (J_\eta(e_{s+t}), J_\eta(e_s)) = (e_{s+t}, e_s)$$
$$= \int_{-\infty}^{\infty} e_{s+t}(x)\,\overline{e_s(x)}\,m_\eta(dx) = \int_{-\infty}^{\infty} e^{itx} m_\eta(dx),$$

where $e_t(x) = e^{itx}$, $-\infty < x < \infty$, from which it follows that the mapping $R^1 \ni t \to \eta_t \in H$ is stationary.

C. It remains to verify formula (4). Since

$$(P(D)R)(t) = \int_{-\infty}^{\infty} e^{itx} P(ix)\,m(dx),$$

$$(P(D)\cdot\overline{P}(-D)R)(t) = (\overline{P}(-D)\cdot P(D)R)(t)$$
$$= \int_{-\infty}^{\infty} e^{itx}\overline{P}(-ix)P(ix)\,m(dx),$$

it remains to note that

$$R_\eta(t) = \int_{-\infty}^{\infty} e^{itx}|P(ix)|^2\,m(dx), \quad \overline{P}(-ix) = \overline{P(ix)}.$$

Example. Let $\{\eta_t: -\infty < t < \infty\}$ be a stationary process with values in a Hilbert space H, whose spectral representation is of the form

$$\eta_t = \int_{-\infty}^{\infty} e^{itx}\mu_\eta(dx), \quad -\infty < t < \infty.$$

Then the differential equation

$$P(D)\,\xi = \eta$$

has a solution $\xi : R^1 \to H$ in the class of stationary processes with bounded spectral measure if and only if

$$\int_{-\infty}^{\infty} |P(ix)|^{-2}\,m_\eta(dx) < \infty,$$

the solution being found in the form

$$\xi_t = \int_{-\infty}^{\infty} e^{itx} P(ix)^{-1}\,\mu_\eta(dx),$$

where m_η is the spectral measure of the process $\{\eta_t\}$. If $P(0) = 0$, then a constant element $h \in H$ orthogonal to $\zeta(t) = \mu(-\infty, t)$ for all t can be added to the solution of the indicated form.

The result of this example simply follows from the theorem (by *reductio ad absurdum*).

§ 4. Integration of Stationary Processes

Let $T = (-\infty, \infty)$, H be a Hilbert space, $\xi : T \to H$ a stationary process, and α a finite (complex-valued) measure on the straight line. It is shown in this section that the process

$$\eta_t = \int_{-\infty}^{\infty} \xi_{t+s}\,\alpha(ds), \quad -\infty < t < \infty, \tag{1}$$

is stationary, and its spectral representation and covariance function are found. First of all, however, we describe the integral involved.

Let $(E, \mathcal{F})$ be a measurable space, and $\alpha : \mathcal{F} \to [0, \infty)$ a finite measure. Construct an integral

$$J(\xi) = \int_E \xi(t)\,\alpha(dt)$$

with values in a separable Hilbert space H of a function ξ on E with values in H (with respect to measure α). The above construction of the integral is actually suitable for the case where H is a separable Banach space. By the σ-algebra of Borel sets in the space H, we understand the minimal σ-algebra containing open sets in the space H (under the norm of this space).

We now give a number of definitions.

D1. A **function** $\xi : E \to H$ is said to be **measurable** if $\xi^{-1}(B) \in \mathcal{F}$ for any Borel set $B \subset H$.

D2. We call **simple** a measurable function $\xi : E \to H$ with finite or countable number of values. Such a function can be written in the form

$$\xi = \sum h_k\,\mathbf{1}_{A_k}, \tag{2}$$

where $\{A_k\}$ is a finite or countable measurable decomposition of the set E, i.e., $\{A_k\} \subset \mathcal{F}$, and $E = \sum A_k$.

D3. A simple function of form (2) is said to be **integrable** if

$$\sum \|h_k\|\,\alpha(A_k) < \infty,$$

in which case the sequence $\{\sum_1^n \alpha(A_k)\,h_k : n = 1, 2, \ldots\}$ converges to a certain element from H (in the norm of this space). Put

$$J(\xi) = \sum_{k \geqslant 1} \alpha(A_k)\cdot h_k.$$

We call the element $J(\xi) \in H$ the **integral** of (a simple integrable function) ξ **with respect to measure** α.

Lemma. Any measurable function $\xi : E \to H$ is the limit of a uniformly convergent sequence of simple functions $\xi_n : E \to H$, i.e., $\|\xi_n(t) - \xi(t)\| \to 0$ uniformly with respect to $t \in E$.

Proof. We construct such a sequence $\{\xi_n\}$ in the following way. Let $\{h_1, h_2, \ldots\}$ be a set everywhere dense in H. For any $\varepsilon = \frac{1}{n} > 0$,

we put

$$S_k = \{h \in H : \|h - h_k\| < \varepsilon\}, \quad \bigcup_{k \geqslant 1} S_k = H,$$

$$B_1 = S_1, \quad B_2 = S_2 S_1^c, \ \ldots, \ B_k = S_k S_1^c \ldots S_{k-1}^c, \ B_k \subset S_k,$$

$$A_k = \xi^{-1}(B_k).$$

Then $\{B_k\}$ and $\{A_k\}$ are measurable decompositions of H and E, respectively (some of the sets may be empty). The function

$$\xi_n = \Sigma 1_{A_k} \cdot h_k$$

is simple, and

$$\|\xi_n(t) - \xi(t)\| < \varepsilon = \frac{1}{n} \quad \forall t \in E.$$

D4. We call a function $\xi: E \to H$ **integrable** if there exists a sequence $\{\xi_n\}$ of simple integrable functions which are uniformly convergent to ξ.

Then the function ξ is measurable, and the sequence $\{J(\xi_n)\} \subset H$ converges (in the norm of the space H) to a certain element $J(\xi) \in H$. We call this element

$$J(\xi) = \int_E \xi(t)\, \alpha(dt) = \int_E \xi\, d\alpha$$

the integral of ξ (with respect to α). It is shown by the usual methods that this definition is correct.

Theorem. Process (1) is stationary. Meanwhile, if

$$R(t) = \int_{-\infty}^{\infty} e^{itx} m(dx), \quad \xi_t = \int_{-\infty}^{\infty} e^{itx} \mu(dx), \quad -\infty < t < \infty$$

are the covariance function and spectral representation of the initial process respectively, where m is a finite (spectral) measure, then

$$R_\eta(t) = \int_{-\infty}^{\infty} e^{itx} m_\eta(dx), \quad \eta_t = \int_{-\infty}^{\infty} e^{itx} \mu_\eta(dx), \quad -\infty < t < \infty$$

are, respectively, the covariance function and spectral representation of process (1), while measures m_η and μ_η are defined by the equalities

$$\mu_\eta(A) = \int_{-\infty}^{\infty} [f(x) \cdot 1_A(x)]\, \mu(dx),$$

$$m_\eta(A) = \|\mu_\eta(A)\|^2 = \int_A |f(x)|^2\, m(dx), \quad A \in \mathbf{B},$$

$$f(x) = \int_{-\infty}^{\infty} e^{ixs} \alpha(ds), \quad -\infty < x < \infty.$$

Corollary. If we take $\alpha = \Sigma c_k \alpha_k$ with a finite number of addends, e.g.,

$$\alpha = (\alpha_1 - \alpha_2) + i(\alpha_3 - \alpha_4),$$

where c_k are complex numbers, $\alpha_k: \mathbf{B} \to [0, \infty)$ finite measures, and define

$$\int_{-\infty}^{\infty} g(x)\, \alpha(dx) = \sum c_k \int_{-\infty}^{\infty} g(x)\, \alpha_k(dx),$$

then the theorem remains valid.

Proof. Taking into account the result of Example 1, § 6, Ch. 6, we have

$$\eta_t = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} e^{i(t+s)x} \mu(dx) \right] \alpha(ds)$$

$$= \int_{-\infty}^{\infty} \left[e^{itx} \int_{-\infty}^{\infty} e^{isx} \alpha(ds) \right] \mu(dx)$$

$$= \int_{-\infty}^{\infty} e^{itx} f(x)\, \mu(dx) = \int_{-\infty}^{\infty} e^{itx} \mu_\eta(dx)$$

(the validity of change of the order of integration in the case of a continuous, bounded function $f(s, x) = e^{i(t+s)x}$, $(s, x) \in R^1 \times R^1$, is left to the reader; see the Remark at the end of the Sec.), while the measure $\mu_\eta: \mathbf{B} \to H$ is orthogonal, σ-additive and finite. Hence,

$$R_\eta(l) = (\eta_{s+t}, \eta_s) = \int_{-\infty}^{\infty} e^{i(s+t)x} f(x) \cdot \overline{e^{isx} f(x)}\, m(dx)$$

$$= \int_{-\infty}^{\infty} e^{itx} |f(x)|^2\, m(dx) = \int_{-\infty}^{\infty} e^{itx} m_\eta(dx);$$

in particular, this means that process (1) is stationary.

Remark. 1°. A measurable function $\xi: E \to H$ is integrable if and only if the function $\|\xi(t)\|$, $t \in E$, is integrable (the integral $\int_E \|\xi(t)\|\, \alpha(dt) < \infty$ is therefore defined). Note that then

$$\left\| \int_E \xi(t)\, \alpha(dt) \right\| \leqslant \int_E \|\xi(t)\|\, \alpha(dt).$$

In fact, for simple functions of form (2), this statement follows from the equality

$$\sum_{k \geqslant 1} \|h_k\|\, \alpha(A_k) = \int_E \|\xi(t)\|\, \alpha(dt)$$

(the convergence of the series is equivalent to the integrability of a simple function).

If a function ξ is integrable, and $\{\xi_n\}$ such a sequence of simple integrable functions that is uniformly convergent to ξ, then, for $\varepsilon = 1$, there exists n_0 such that $\| \xi(t) - \xi_{n_0}(t) \| < \varepsilon \; \forall \, t \in E$; hence, the measurable function $\| \xi(t) \|$, $t \in E$, is majorized by the integrable function $\| \xi_{n_0}(t) \| + 1$, $t \in E$, which implies the integrability of the function $\| \xi(t) \|$, $t \in E$.

Conversely, let the function $\| \xi(t) \|$, $t \in E$, be integrable for a measurable function ξ. By the lemma, there exists such a sequence $\{\xi_n\}$ of simple functions that is uniformly convergent to ξ. It suffices to show that all these simple functions, except, possibly, a finite number of them, are integrable. Indeed, for $\varepsilon = 1$, there exists n_0 such that

$$\| \xi_n(t) - \xi(t) \| \leqslant 1 \text{ for } t \in E, \; n \geqslant n_0,$$

where

$$\| \xi_n(t) \| \leqslant \| \xi(t) \| + 1 \text{ for } t \in E, \; n \geqslant n_0;$$

therefore, the measurable functions $\| \xi_n(t) \|$, $t \in E$, $n \geqslant n_0$ are integrable as majorizable by an integrable function. However, as has already been shown, if we require that ξ_n should be integrable, the integrability of the function $\| \xi_n(t) \|$, $t \in E$, is sufficient (and also necessary).

2°. If $\xi : E \to H$ is an integrable function, $c \in H$, then $E \ni t \to (c, \xi(t)) \in R^1$ is integrable, and

$$(3) \qquad \left(c, \int_E \xi \, d\alpha\right) = \int_E (c, \xi) \, d\alpha.$$

This equality indicates the following method of defining the integral $J(\xi) = \int_E \xi d\alpha$. If ξ is integrable, then the functional $H \ni c \to f(c) = \int_E (c, \xi) \, d\alpha$ is additive and bounded, since

$$|f(c)| \leqslant \| c \| \cdot \int_E \| \xi(t) \| \, \alpha(dt);$$

therefore, by the Riesz theorem, there exists an element $J(\xi) \in H$ such that $f(c) = (c, J(\xi)) \; \forall \, c \in H$. It is natural to define $J(\xi)$ just as the integral of ξ with respect to α.

By Item 1°, it is convenient in this definition to call the measurable function ξ integrable if $\| \xi(t) \|$, $t \in E$, is integrable, from which it follows, in particular, that $t \to (c, \xi(t))$ is integrable, because it is measurable and majorized by an integrable function, viz.,

$$| (c, \xi(t)) | \leqslant \| c \| \cdot \| \xi(t) \|.$$

We now prove (3). As the relation is obvious for simple integrable functions, it remains to show that if $\{\xi_n\}$ is a sequence of simple integrable functions, which is uniformly convergent to ξ, then

$$\left(c, \int_E \xi_n \, d\alpha\right) \to \left(c, \int_E \xi \, d\alpha\right),$$

and

$$\int_E (c, \xi_n) \, d\alpha \to \int_E (c, \xi) \, d\alpha$$

due to

$$\left| \left(c, \int_E (\xi_n - \xi) \, d\alpha\right) \right| \leqslant \| c \| \cdot \sup_t \| \xi_n(t) - \xi(t) \| \cdot \alpha(E) \to 0,$$

and

$$\left| \int_E (c, \xi_n - \xi) \, d\alpha \right| \leqslant \| c \| \cdot \sup_t \| \xi_n(t) - \xi(t) \| \cdot \alpha(E) \to 0.$$

3°. A stationary process $\xi : R^1 \to H$ with continuous covariance function R (we have always considered just this case if ξ is a continuous-time process) is integrable. In fact, the measurability of the function ξ follows from its continuity, viz.,

$$\| \xi_{t+h} - \xi_t \|^2 = 2R(0) - R(h) - R(-h), \; R(-h) = \overline{R(h)}.$$

In addition, since $\| \xi_t \|^2 = R(0)$, the function $t \to \| \xi_t \|$ is integrable; therefore, by the statement of Item 1°, ξ is integrable.

4°. We now state the possibility of changing the order of integration (the Fubini theorem) in the following situation. Consider two measurable spaces $(E_1, \mathcal{F}_1)$ and $(E_2, \mathcal{F}_2)$, a finite measure $\alpha : \mathcal{F}_1 \to [0, \infty)$, and orthogonal, σ-additive, finite measure $\mu : \mathcal{F}_2 \to H$.

A finite (structural) measure $m : \mathcal{F}_2 \to [0, \infty)$ is defined in the usual way by the equality $m(A) = \| \mu(A) \|^2$, $A \in \mathcal{F}_2$ (see § 6, Ch. 6) by means of the measure μ. Put

$$E = E_1 \times E_2, \quad \mathcal{F} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2), \quad \lambda = \alpha \times m.$$

We now prove the following statement. If a (complex-valued) function $f \in L^2_\lambda = L^2(E, \mathcal{F}, \lambda)$ and measures α, m are complete, then

$$(4) \qquad \int_{E_1} \left[\int_{E_2} f \, d\mu \right] d\alpha = \int_{E_2} \left[\int_{E_1} f \, d\alpha \right] d\mu,$$

where all integrals involved make sense. In particular, this formula is valid if f is bounded, measurable and, therefore, integrable (if raised to any positive power).

(1) Formula (4) is obvious for functions of the form $f = 1_A$, where $A \in \mathcal{F}_1 \times \mathcal{F}_2$, and, therefore, for finite linear combinations of such functions; hence, also for functions of the form $f = \Sigma c_k 1_{A_k}$

with a finite number of addends, but where A_k belongs to $\mathscr{F}_0$, the minimal algebra containing $\mathscr{F}_1 \times \mathscr{F}_2$. Such functions are said to be **elementary**. We show that the set of elementary functions is everywhere dense in L^2_λ. Since the set of functions of the form $f = \Sigma c_k 1_{A_k}$, $A_k \in \mathscr{F} \supset \mathscr{F}_0$, with a finite number of addends, is everywhere dense in L^2_λ, it suffices to show that, for $\varepsilon > 0$ and $A \in \mathscr{F}$, there exists such $B \in \mathscr{F}_0$ that $\| 1_B - 1_A \| < \varepsilon$ under the norm of L^2_λ.

In fact, according to the construction of the extension of measure λ from $\mathscr{F}_0$ to $\mathscr{F}$, we have

$$\lambda(A) = \inf_{\substack{\bigcup_n B_n \supset A \\ \{B_n\} \subset \mathscr{F}_0}} \sum \lambda(B_n),$$

where we can assume that the sets $\{B_n\}$ are pairwise disjoint. For $\delta > 0$, we select such $\{B_n\} \subset \mathscr{F}_0$ that $\sum_{n \geqslant 1} B_n \supset A$, and $\lambda(A) \geqslant \sum_{n \geqslant 1} \lambda(B_n) - \delta$, and then choose N so that $\sum_{n > N} \lambda(B_n) < \delta$. Put $B = \sum_1^N B_n$, $C = \sum_{n > N} B_n$. We have

$$B \in \mathscr{F}_0,\ B + C \supset A,\ \lambda(B) - \lambda(A) < \delta,\ \lambda(C) < \delta.$$

Since

$$\| 1_B - 1_A \|^2 = \int_E |1_B - 1_A|^2 \, d\lambda,$$

and

$$|1_B - 1_A|^2 = |1_B - 1_A| \leqslant [1_B + 1_C - 1_A] + 1_C = 1_B - 1_A + 2 \cdot 1_C,$$

we obtain

$$\| 1_B - 1_A \|^2 \leqslant \lambda(B) - \lambda(A) + 2\lambda(C) < 3\delta.$$

It remains to put $3\delta = \varepsilon^2$.

(2) We now show that the four integrals involved in formula (4) make sense. Define two functions f_x and f_y on E_2 and on E_1, respectively, by putting

$$f_x(y) = f(x, y), \quad f_y(x) = f(x, y).$$

Since $f \in L^2_\lambda$, we get

$$f_x \in L^2(E_2, \mathscr{F}_2, m) = L^2_m$$

(almost for all x with respect to measure α), and $f_y \in L^2(E_1, \mathscr{F}_1, \alpha) = L^2_\alpha$; then $f_y \in L^1(E_1, \mathscr{F}_1, \alpha) = L^1_\alpha$ (almost for all y with respect to measure m). Now,

(a) the integral $\xi(x) = \int_{E_2} f d\mu = \int_{E_2} f_x d\mu$ is defined, since $f_x \in L^2_m$,

(b) the integral $g(y) = \int_{E_1} f d\alpha = \int_{E_1} f_y d\alpha$ is defined, because $f_y \in L^1_\alpha$,

(c) the integral $\int_{E_1} \xi d\alpha$ is defined, for the function $x \to \| \xi(x) \|^2 = \int_{E_2} |f_x|^2 \, dm$ is integrable with respect to measure α; therefore, the function $x \to \| \xi(x) \|$ is integrable with respect to measure α, which implies (by 1°) the integrability of ξ, and

(d) the integral $\int_{E_2} g d\mu$ is defined, since $g \in L^2_m$, which follows from the inequality

$$|g(y)|^2 \leqslant \Big[\int_{E_1} |f_y| \, d\alpha\Big]^2 \leqslant \alpha(E_1) \int_{E_1} |f_y|^2 \, d\alpha, \quad y \in E_2,$$

and integrability of the right-hand side with respect to measure m.

(3) Now, let $f \in L^2_\lambda$, and $\{f_n\}$ be a sequence of elementary functions (for which formula (4) has already been proved), which is convergent to f in the norm of the space L^2_λ. It now suffices to show that

$$\alpha_n = \Big\| \int_{E_1} \Big[\int_{E_2} (f_n - f) \, d\mu \Big] d\alpha \Big\| \to 0,$$

and

$$\beta_n = \Big\| \int_{E_2} \Big[\int_{E_1} (f_n - f) \, d\alpha \Big] d\mu \Big\| \to 0.$$

Define a function $\xi : E_1 \to H$ by putting $\xi(x) = \int_{E_2} f_x d\mu$. Similarly, we define ξ_n in terms of f_n. Then

$$\alpha_n^2 = \Big\| \int_{E_1} (\xi_n - \xi) \, d\alpha \Big\|^2 \leqslant \Big[\int_{E_1} \| \xi_n - \xi \| \, d\alpha \Big]^2$$

$$\leqslant \alpha(E_1) \int_{E_1} \| \xi_n - \xi \|^2 \, d\alpha = \alpha(E_1) \int_{E_1} \Big[\int_{E_2} |f_n - f|^2 \, dm \Big] d\alpha$$

$$= \alpha(E_1) \int_E |f_n - f|^2 \, d\lambda = \alpha(E_1) \| f_n - f \|^2,$$

$$\beta_n^2 = \int_{E_2} \Big| \int_{E_1} (f_n - f) \, d\alpha \Big|^2 dm$$

$$\leqslant \alpha(E_1) \int_{E_2} \Big[\int_{E_1} |f_n - f|^2 \, d\alpha \Big] dm = \alpha(E_1) \| f_n - f \|^2.$$

Supplement: Regenerative Processes with Dependent Regeneration Cycles of Markov Type

§ 1. Semi-Markov Renewal Process

Consider a semi-Markov process $\zeta = \{\zeta(t), t \geqslant 0\}$ with a discrete state space E. Such a process is given by a sequence of RV's (ζ_n, z_n), $n = 0, 1, 2, \ldots$, with values in $E \times R_+^1$, where $R_+^1 = [0, \infty)$. The RV's obey a Markov dependence of the form

$$P(\zeta_{n+1} = j,\ z_{n+1} < y \mid \zeta_n = i,\ z_n = x) = P(\zeta_{n+1} = j,\ z_{n+1} < y \mid \zeta_n = i),$$

which is homogeneous in time. In this case, the sequence of RV's $\zeta_0, \zeta_1, \ldots$ forms a homogeneous Markov chain with the state space E. Assuming that the transition from ζ_n to ζ_{n+1} is carried out in time z_{n+1} (and, thereby, z_{n+1} is the sojourn at the state ζ_n), we denote the state at an epoch $t \geqslant 0$ by $\zeta(t)$. The process $\zeta = \{\zeta(t), t \geqslant 0\}$ so defined is called a **semi-Markov process** (SMP).

An SMP ζ is defined by a set of functions

$$F_{ij}(x) = P(\zeta_{n+1} = j,\quad z_{n+1} < x \mid \zeta_n = i),\quad x \geqslant 0$$

$i, j \in E$, up to stochastic equivalence. Put $F_{ij}(+\infty) = p_{ij}$; if $p_{ij} \neq 0$, then let $A_{ij}(x) = F_{ij}(x)/F_{ij}(+\infty)$. Note that the "embedded" Markov chain $\{\zeta_n, n \geqslant 0\}$ is determined by a transition probability matrix $P = \{p_{ij}\}$, but if $p_{ij} \neq 0$, then A_{ij} is the DF of the sojourn of an SMP at a state i in transition to a state j. The initial state of an SMP is specified by the distribution $\{p_i\}$ of the RV ζ_0. We assume that $z_0 = 0$.

A natural generalization of a recurrent renewal process is given by a process $\bar{\nu} = \{\bar{\nu}(t), t \geqslant 0\}$, where $\bar{\nu}(t)$ is the number of changes of the SMP ζ in the interval $[0, t)$, in which case

$$\bar{\nu}(t) = \sup\{n : t_n < t\},$$

$$t_n = z_1 + \ldots + z_n,\quad n \geqslant 1,\quad t_0 = 0.$$

If $\zeta_0 = i$, denote the number of transitions of the SMP ζ to a state j in the interval $[0, t)$ by $\nu_{ij}(t)$. Note that $\nu_{ij}(t)$ is a recurrent delayed renewal process. Denote the collection of such processes by $\nu(t) = \{\nu_{ij}(t) : i, j \in E\}$, $t \geqslant 0$, and call it a **semi-Markov renewal process** (SMRP) **generated by** the SMP ζ.

§ 2. Sufficient Condition for SMRP Moment Existence

We assume that the following assumptions are valid.

A1. An embedded chain with transition probability matrix $P = \{p_{ij};\ i, j \in E\}$ is irreducible and recurrent.

A2. There exists such a pair of states i_0 and j_0 from E that

$$p_{i_0 j_0} > 0,\ \text{and}\ A_{i_0 j_0}(+0) < 1.$$

Remark. A state i of a chain $\{\zeta_n : n \geqslant 0\}$ is said to be **recurrent** if, starting from i, the probability of return to i equals unity. Formally, it means that

$$f_{ii}^* = P\{\bigcup_{n \geqslant 1} (\zeta_n = i) \mid \zeta_0 = i\} = 1.$$

For an irreducible chain, all the states are recurrent if at least one is recurrent.

Theorem. There exists such a sequence of positive numbers θ_j, $j \in E$, that

$$E e^{\theta_j \nu_{ij}(t)} < +\infty$$

for all i and j from E. In particular, this implies the existence of moments of all orders of the RV $\nu_{ij}(t)$.

Proof. Denote by τ_{ij} the time between transitions to states i and j in the indicated sequence. Assuming that i and j are fixed, we put

$$P(\tau_{ij} < x) = A_1(x),\quad P(\tau_{jj} < x) = A(x).$$

Then $\nu_{ij}(t)$, $t \geqslant 0$, is a recurrent delayed renewal process, which is determined by distributions A_1 and A (note that the distributions A_1 and A are non-degenerate). For such a renewal process, the condition $A(+0) < 1$ provides for the existence of a positive number $\theta\ (= \theta_j)$ such that

$$E \exp\{\theta \nu_{ij}(t)\} < \infty,\quad t > 0.$$

It remains to verify that $A(+0) < 1$ follows from A1, A2.

Denote by $\Pi = \{\pi\}$ the collection of ordered sequences of such states of the form $\pi = (s_0, s_1, \ldots, s_n)$, $s_0 = s_n = j$, $s_k \neq j$, for $0 < k < n$, that

$$p_\pi = p_{s_0 s_1} \cdot p_{s_1 s_2} \cdot \ldots \cdot p_{s_{n-1} s_n} > 0.$$

Moreover, if we put

$$A_\pi = A_{s_0 s_1} * A_{s_1 s_2} * \ldots * A_{s_{n-1} s_n},$$

where $*$ means the convolution operation, then

$$A(x) = P(\tau_{ij} < x) = \sum_{\pi \in \Pi} p_\pi A_\pi(x).$$

Due to A1 and A2, there exists an ordered sequence $\pi_0 \in \Pi$ containing the states i_0 and j_0 as neighbouring. Since

$$\sum_{\pi \in \Pi} p_\pi = 1, \quad A_{\pi_0}(x) \leqslant A_{i_0 j_0}(x)$$

(for any two distributions A and B on $[0, \infty)$, we have $(A * B)(x) \leqslant A(x)$),

$$A(x) \leqslant 1 - p_{\pi_0} + p_{\pi_0} A_{i_0 j_0}(x),$$

which gives $A(+0) < 1$ because of A2.

§ 3. Elementary SMRP Renewal Theorem

Consider an SMRP $\nu(t) = \{\nu_{ij}(t)\}$, $t \geqslant 0$, and assume that A1 and A2, § 2 are valid, thus providing, in particular, for the existence of the first moments

$$H_{ij}(t) = E\nu_{ij}(t).$$

We call $H(t) = \{H_{ij}(t)\}$, $t \geqslant 0$, the **renewal function of the SMRP.** Note that $H_{ij}(t)$ is the average number of transitions of the SMP ζ to a state j in an interval $[0, t)$, proceeding from an initial state i.

In addition, we put

$$w_{ij} = \int_0^\infty x\, dF_{ij}(x).$$

Note that $w_{ij} = 0$ when $p_{ij} = 0$, and if $p_{ij} > 0$, then

$$w_{ij} = p_{ij} \int_0^\infty x\, dA_{ij}(x) = p_{ij} u_{ij}.$$

Theorem.

(1) $$\lim_{t \to \infty} \frac{H_{ij}(t)}{t} = a_j,$$

where

$$a_j = \frac{u_j}{\sum_\alpha u_\alpha w_\alpha}, \quad w_\alpha = \sum_\beta w_{\alpha\beta},$$

$\{u_\alpha, \alpha \in E\}$ being a non-zero and non-negative solution of the system

(2) $$u_\beta = \sum_\alpha u_\alpha p_{\alpha\beta}, \quad \beta \in E.$$

Remark. 1°. The value w_α has the meaning of the average sojourn at a state α (up to the transition to the next state).

2°. A non-zero and non-negative solution of system (2) exists. In fact, by Theorem 9.7, Part 1, of *Markov chains with stationary transition probabilities* by K.-L. Chung (Springer, New York, 1967), such a solution is of the form $u_i = c \cdot \varepsilon_{ij}$, where

$$\varepsilon_{ij} = \lim_{N \to \infty} \frac{1 + \sum_1^N p_{ii}^n}{1 + \sum_1^N p_{jj}^n},$$

the assumption of irreducibility of the embedded chain providing for the existence of this limit, and $0 < \varepsilon_{ij} < \infty$ for all $i, j \in E$.

Corollary. Let there exist the stationary distribution $\{\pi_i\}$ of the embedded chain, and $\pi_i \geqslant 0$, $\Sigma \pi_i = 1$. It then follows from A1 that $\pi_i > 0$ for all $i \in E$, in which case

(3) $$a_j = a\pi_j, \quad a^{-1} = \sum_k \pi_k w_k,$$

where w_k is the average sojourn at a state k.

Proof of the theorem. We make use of the elementary renewal theorem for a recurrent delayed renewal process which is determined by two DF's A_1 and A. By this theorem,

$$\lim_{t \to \infty} \frac{H(t)}{t} = \hat{a},$$

where $H(t)$ is a renewal function, and $\hat{a}^{-1} = \int_0^\infty x dA(x)$, while assuming that $\hat{a} = 0$ if $\int_0^\infty x dA(x) = \infty$. (This theorem is also valid in the case where A_1 and A are improper DF's; meanwhile, if $A(+\infty) < 1$, then $\hat{a} = 0$.)

In our case, $\nu_{ij}(t)$, $t \geqslant 0$, is a recurrent delayed renewal process determined by the functions

$$A_1(x) = P(\tau_{ij} < x), \quad A(x) = P(\tau_{jj} < x).$$

It follows from A1 that $A_1(+\infty) = f_{ij}^* = 1$, $A(+\infty) = f_{jj}^* = 1$, where f_{ij}^* is the probability to enter a state j (at least once), proceeding from a state i, i.e.,

$$f_{ij}^* = P\{\bigcup_{n \geqslant 1} (\zeta_n = j) | \zeta_0 = i\}.$$

Therefore, it remains to find only

$$a_j^{-1} = E\tau_{jj} = \int_0^\infty x dA(x).$$

We introduce necessary notation. For the embedded Markov chain $\{\xi_n;\ n \geqslant 0\}$, we put

$$p^n_{\alpha\beta}(\gamma) = P(\xi_n = \beta,\ \xi_k \neq \gamma \quad \text{for} \quad 0 < k < n \mid \xi_0 = \alpha),\ n \geqslant 1$$

i.e., $p^n_{\alpha\beta}(\gamma)$ is the probability of transition from α to β in n steps without entering a state γ during the intermediate steps, viz.,

$$p_{\alpha\beta}(\gamma) = \sum_{n\geqslant 1} p^n_{\alpha\beta}(\gamma) \tag{4}$$

(if the series diverges, then its value is $+\infty$). It is obvious that $p_{\alpha\beta}(\gamma)$ is the ME of the number of entrances to a state β until the Markov chain starting with a state α enters a state γ.

Then

$$a_j^{-1} = E\tau_{jj} = w_{jj} + \sum_{n\geqslant 1}\sum_{\alpha\neq j}(w_{j\alpha}p^n_{\alpha j}(j) + p^n_{j\alpha}(j)\,w_{\alpha j}) + \sum_{n\geqslant 3}\sum_{k=1}^{n-2}\sum_{\alpha\neq j}\sum_{\beta\neq j} p^k_{j\alpha}(j)\,w_{\alpha\beta}p^{n-1-k}_{\beta j}(j),$$

whence, by changing the order of integration, employing notation (4), and replacing the summation variable by the formula $n - 1 - k = m$, we obtain

$$\begin{aligned} a_j^{-1} &= w_{jj} + \sum_{\alpha\neq j}\sum_{n\geqslant 1}[w_{j\alpha}p^n_{\alpha j}(j) + p^n_{j\alpha}(j)\,w_{\alpha j}] \\ &\quad + \sum_{\alpha\neq j}\sum_{\beta\neq j}\sum_{k\geqslant 1}\sum_{m\geqslant 1} p^k_{j\alpha}(j)\,w_{\alpha\beta}p^m_{\beta j}(j) \\ &= w_{jj} + \sum_{\alpha\neq j}[w_{j\alpha}p_{\alpha j}(j) + p_{j\alpha}(j)\,w_{\alpha j}] \\ &\quad + \sum_{\alpha\neq j}\sum_{\beta\neq j} p_{j\alpha}(j)\,p_{\beta j}(j)\,w_{\alpha\beta}. \end{aligned}$$

It remains to make use of the formulas

$$p_{\alpha j}(j) = f^*_{\alpha j},\quad p_{\beta j}(j) = f^*_{\beta j},$$
$$p_{j\alpha}(j) = f^*_{j\alpha}\cdot\varepsilon_{\alpha j}.$$

The first two of them are obvious, whereas the latter follows from formula (9.15), P. 1, in the book by K.-L. Chung if we put $H = \varnothing$, which gives

$$\varepsilon_{\alpha j} = \lim_{N\to\infty}\frac{1 + \sum_{n=1}^{N} p^n_{\alpha\alpha}}{1 + \sum_{n=1}^{N} p^n_{jj}} = \frac{p_{j\alpha}(j)}{p_{j\alpha}(\alpha)} = \frac{p_{j\alpha}(j)}{f^*_{j\alpha}}.$$

In view of A1, we have $f^*_{\alpha\beta} = 1$ for all α and $\beta \in E$. Now,

$$\begin{aligned} a_j^{-1} &= w_{jj} + \sum_{\alpha\neq j}(w_{j\alpha} + \varepsilon_{\alpha j}w_{\alpha j}) + \sum_{\alpha\neq j}\sum_{\beta\neq j}\varepsilon_{\alpha j}w_{\alpha\beta} \\ &= \sum_{\alpha} w_{j\alpha} + \sum_{\alpha\neq j}\varepsilon_{\alpha j}\sum_{\beta} w_{\alpha\beta} = w_j + \sum_{\alpha\neq j}\varepsilon_{\alpha j}w_\alpha = \sum_{\alpha}\varepsilon_{\alpha j}w_\alpha \end{aligned}$$

if we take into account that $\varepsilon_{jj} = 1$. In this case, since $\varepsilon_{\alpha j} = \varepsilon_{\alpha j_0}\varepsilon_{j_0 j}$, we have

$$a_j^{-1} = \sum_{\alpha}\varepsilon_{\alpha j_0}w_\alpha\varepsilon_{j_0 j} = a_{j_0}^{-1}\varepsilon_{j_0 j},$$

whence

$$a_j = \frac{\varepsilon_{jj_0}}{\sum_{\alpha}\varepsilon_{\alpha j_0}w_\alpha}$$

due to the equality $\varepsilon_{j_0 j} = \varepsilon^{-1}_{jj_0}$, thus yielding (1) if we make use of Remark 2°.

§ 4. SMRP Blackwell Theorem

In accordance with the terminology of the book by W. Feller, V. 2, a distribution on R^1 is said to be **arithmetic** if its points of increase are contained in a set of the form $\{nd : n \text{ is an integer}\}$, $d \neq 0$. If, however, the points of increase are of the form $c + nd$, $n = 0, \pm 1, \pm 2, \ldots$, then the distribution is said to be **lattice**.

We call a set of distributions $\{A_\alpha\}$ on R^1 arithmetic (lattice) with **lattice constant** $d > 0$ if the points of increase of each of the distributions are of the form nd (resp. $c + nd$) for $n = 0, \pm 1, \pm 2, \ldots$.

We call a set of distributions on R^1 arithmetic (lattice) if it is arithmetic (lattice) with a certain positive lattice constant. Otherwise, such a set is termed **non-arithmetic (non-lattice)**. Together with A1 and A2, we consider the following.

A3. At least one of the following propositions is true, viz.,

1. A set of distributions $\{A_{ij}:\ p_{ij} > 0\}$ is non-arithmetic, and the point 0 is a point of increase of each of the distributions A_{ij} when $p_{ij} > 0$.
2. There exists a pair of states i_1 and j_1 such that $p_{i_1 j_1} > 0$, and the distribution $A_{i_1 j_1}$ is non-lattice.
3. For any state j, there exists an ordered set of states $\pi = (s_0, s_1, \ldots, s_n)$ such that $s_0 = s_n = j$, $s_k \neq j$ for $0 < k < n$, $p_\pi > 0$, and the distribution A_π contains two rationally incommensurable points of increase (two points x and y are said to be rationally commensurable if $mx = ny$ for certain integers m and n).

Theorem. Let A1-A3 be valid. Then

$$H_{ij}(t + h) - H_{ij}(t) \to a_j h \tag{1}$$

as $t \to \infty$ for any $h > 0$ and any pair of states i, j, whereas the number a_j has been defined in the previous section.

Proof. Note first of all that A1 provides for the equality $P(\tau_{ij} < \infty) = 1$ for all i and j. By the Blackwell theorem on recurrent delayed renewal processes, it suffices to verify that the distribution

$$A(x) = P(\tau_{jj} < x), \quad x > 0$$

is non-arithmetic.

Just like in § 2, we denote by $\Pi = \{\pi\}$ the collection of ordered sequences of states of the form

$$\pi = (s_0, s_1, \ldots, s_n), \quad s_0 = s_n = j, \quad s_k \neq j \text{ for } 0 < k < n$$

such that

$$p_\pi = p_{s_0 s_1} \cdot \ldots \cdot p_{s_{n-1} s_n} > 0. \tag{2}$$

Moreover, if we put

$$A_\pi = A_{s_0 s_1} * \ldots * A_{s_{n-1} s_n}, \tag{3}$$

where $*$ denotes the convolution operation, then

$$A(x) = \sum_{\pi \in \Pi} p_\pi A_\pi(x), \quad x > 0. \tag{4}$$

We make use of the following simple statements.

S1. Let $F = \sum p_k F_k$, where $p_k > 0$, $\sum p_k = 1$, F and F_k are distributions on R^1. Then the distribution F is lattice if (and only if) each of the distributions F_k is lattice with the same lattice constant.

S2. Let distributions F, F_1 and F_2 on R^1 be related by the equality $F = F_1 * F_2$. Then the distribution F is lattice if (and only if) each of the distributions F_1 and F_2 is lattice (with the same lattice constant).

It follows from these statements and (2)-(4) that a distribution A is lattice if (and only if) the set of distributions $\{A_{ij} : p_{ij} > 0\}$ is lattice, with the effect that Items 2 and 3 in A3 (along with A1 and A2) both imply that the distribution A is non-arithmetic.

It remains to verify that Item 1 in A3, together with A1 and A2, provide for the validity of the conclusion of the theorem. In fact, S1 and S2 remain valid if, for each of the distributions, F, F_k, the point 0 is a point of increase, and the term "lattice" is replaced by "arithmetic". That $\{A_{ij} : p_{ij} > 0\}$ is arithmetic then follows from the arithmeticity of A, which contradicts the conditions of the theorem. Therefore, the distribution A is non-arithmetic.

§ 5. SMRP Key Theorem of Renewal Theory

Theorem. Let Q be a directly Riemann integrable function on $[0, \infty)$, and A1-A3 valid. Then

$$\lim_{t \to \infty} \int_0^t Q(t-x)\, dH_{ij}(x) = a_j \int_0^\infty Q(x)\, dx.$$

Note that the integral on the left-hand side is understood in the sense of Lebesgue-Stieltjes. It exists, since Q is a bounded and Borel measurable function, whereas H_{ij} monotonic. The integral on the right-hand side is usual Riemann.

This statement coincides with the statement of Example 3, § 4, Ch. 4, in which we should put

$$H_1 = H_{ij}, \quad H = H_{jj},$$
$$A_1(x) = P(\tau_{ij} < x), \quad A(x) = P(\tau_{jj} < x),$$
$$A_1(\infty) = f_{ij}^* = 1, \quad A(\infty) = f_{jj}^* = 1.$$

Besides, we have seen in the previous section that A is a non-arithmetic distribution.

§ 6. Limit Theorem for Regenerative Processes with Dependent Regeneration Cycles of Markov Type

1°. We call a **cycle** of duration z an ordered pair (z, ξ), where z is a non-negative real (random) number, and ξ a random function defined on $[0, \infty)$, which takes values in a measurable space $(X, \mathbf{B})$. We assume that the function $(t, w) \to \xi(t, w)$ is measurable.

Consider a sequence of cycles $\{(z_k, \xi_k) : k \geqslant 1\}$, and form a new process $\xi(t)$, $t \geqslant 0$, by placing intervals $[0, z_1)$, $[0, z_2)$, ... consecutively, one after another, on the half-line $[0, \infty)$, while assuming that ξ coincides with ξ_k on the k-th interval.

Formally, this means that $\xi(t) = \xi_k(t - t_{k-1})$ when $t_{k-1} \leqslant t < t_k$, with $t_0 = 0$, $t_k = z_1 + \ldots + z_k$, $k \geqslant 1$.

Consider another sequence $\{\zeta_n : n \geqslant 0\}$ of RV's which take values from a discrete set E, and assume that the sequence

$$\{(\zeta_n, z_n, \xi_n(t), 0 \leqslant t < z_n)\}_{n \geqslant 0}, \quad z_0 = 0$$

satisfies the following assumption.

A4. The sequence $\{(\zeta_n, z_n, \xi_n) : n \geqslant 0\}$ obeys a homogeneous Markov dependence, and

$$P\{\zeta_{n+1} = j, z_{n+1} > y, \xi_{n+1}(y) \in B \mid \zeta_n = i, z_n = x\}$$
$$= P\{\zeta_{n+1} = j, z_{n+1} > y, \xi_{n+1}(y) \in B \mid \zeta_n = i\}.$$

Remark. Instead of A4, it suffices to require that

(A4') the sequence $\{(\zeta_n, z_n)\}$ should obey a homogeneous Markov dependence of the form

$$P(\zeta_{n+1} = j, z_{n+1} > y \mid \zeta_n = i, z_n = x) = P(\zeta_{n+1} = j, z_{n+1} > y \mid \zeta_n = i),$$

and

$$P(z_{n+1} > y, \xi_{n+1}(y) \in B \mid \zeta_0 = i_0, \zeta_n = i, z_1 + \ldots + z_n = x)$$
$$= P(z_{n+1} > y, \xi_{n+1}(y) \in B \mid \zeta_n = i)$$

be independent of n, in which case the sequence $\{(\zeta_n, z_n) : n \geqslant 0\}$ generates an SMP $\zeta(t)$, $t \geqslant 0$ (see § 1). The process $\xi(t)$, $t \geqslant 0$, defined above is called a **regenerative process (with dependent regeneration cycles of Markov type).**

The goal of this section is to clarify under which conditions the limit

$$\lim_{t\to\infty} P\{\zeta(t) = j,\ \xi(t) \in B\} \tag{1}$$

exists, and to indicate a method of its evaluation.

Put

$$\mu_j(y) = \mu_j(y, B) = P(z_{n+1} > y,\ \xi_{n+1}(y) \in B \mid \zeta_n = j).$$

The main method of evaluating limit (1) is based on the relation

$$P(\zeta(t) = j,\ \xi(t) \in B \mid \zeta_0 = i) = \delta_{ij} P(z_1 > t,\ \xi(t) \in B \mid \zeta_0 = j)$$

$$+ \sum_{n\geqslant 1} \int_0^t P(z_{n+1} > t-\tau,\ \xi_{n+1}(t-\tau) \in B \mid \zeta_n = j) d_\tau P(t_n < \tau,\ \zeta_n = j \mid \zeta_0 = i)$$

$$= \delta_{ij}\mu_j(t, B) + \sum_{n\geqslant 1} \int_0^t \mu_j(t-\tau, B)\, d_\tau P(t_n < \tau,\ \zeta_n = j \mid \zeta_0 = i)$$

which is derived from the total probability formula, or

$$P(\zeta(t) = j,\ \xi(t) \in B \mid \zeta_0 = i) \tag{2}$$

$$= \delta_{ij}\cdot\mu_j(t, B) + \int_0^t \mu_j(t-\tau, B)\, dH_{ij}(\tau),$$

where

$$H_{ij}(\tau) = \sum_{n\geqslant 1} P(t_n < \tau,\ \zeta_n = j \mid \zeta_0 = i) = E_i \sum_{n\geqslant 1} 1_{\{t_n<\tau,\ \zeta_n = j\}}$$

is the average number of transitions to a state j of the SMP $\zeta(t)$, $t \geqslant 0$, in the interval $[0, \tau)$ (when starting with a state i).

We shall see below how the key SMRP renewal theorem (see § 5) will permit to evaluate relation (1) by means of (2).

We define a set $F^n = \{F^n_{ij}\}$, proceeding from a set $F = \{F_{ij}\}$ of functions

$$F_{ij}(x) = P(\zeta_{n+1} = j,\quad z_{n+1} < x \mid \zeta_n = i),$$

and putting consecutively

$$F^1_{ij} = F_{ij},\quad F^{n+1}_{ij} = \sum_k F_{ik} * F^n_{kj},\quad n \geqslant 1.$$

Note that

$$F^n_{ij}(x) = P(t_n < x,\ \zeta_n = j \mid \zeta_0 = i).$$

Put

$$F^0_{ij}(x) = \delta_{ij}\cdot H(x),\quad H(x) = \begin{cases} 0 \text{ for } x \leqslant 0 \\ 1, \text{ for } x > 0 \end{cases},$$

with

$$H_{ij}(t) = \sum_{n\geqslant 1} F^n_{ij}(t),\quad t \geqslant 0. \tag{3}$$

Theorem. Let A1-A4 and at least one of the following conditions be fulfilled.

C1. The function $\mu_j(t, B)$ is Riemann integrable on each finite interval with respect to t, and $0 < w_j < \infty$.

C2. For any state $i \in E$, there exists an integer $n \geqslant 1$ such that each of the functions F^n_{ik}, $k \in E$, is absolutely continuous, and $0 < w_j < \infty$.

Then

$$\lim_{t\to\infty} P(\xi(t) \in B,\quad \zeta(t) = j \mid \zeta_0 = i) = a_j \int_0^\infty \mu_j(x, B)\, dx \tag{4}$$

for any $i \in E$.

Recall that w_j is the average sojourn at a state j (up to the transition to the next state).

We assume that A1-A4 are fulfilled. The states i and j are regarded as fixed, and are not used in the case of certain notations.

Put

$$A_1(x) = P(\tau_{ij} < x),\ A(x) = P(\tau_{jj} < x),\ x \geqslant 0.$$

Lemma. Let there exist an integer $n \geqslant 0$ such that the function

$$Q(t) = (\mu_j * F)(t) = \int_0^t \mu_j(t-x, B)\, dF(x),\quad t \geqslant 0,$$

where $F = A^{*n}$, is directly Riemann integrable on $[0, \infty)$. Then

$$\lim_{t\to\infty} P(\xi(t) \in B,\ \zeta(t) = j \mid \zeta_0 = i) = a_j \int_0^\infty \mu_j(x, B)\, dx.$$

Proof is based on formula (2) and the expression

$$H_{ij} = A_1 + A_1 * A + A_1 * A^{*2} + \ldots = A_1 + A_1 * \sum_{n\geqslant 1} A^{*n}.$$

If $n = 0$, then $Q(t) = \mu_j(t, B)$, with the statement following from the key renewal theorem (see § 5) and

$$\mu_j(t, B) \leqslant P(z_{n+1} > t \mid \zeta_n = j) \to 0 \text{ as } t \to \infty.$$

The same argument is also applicable when $n \geqslant 1$, since it then follows from (2) that

$$P(\xi(t) \in B, \zeta(t) = j \mid \zeta_0 = i) = \delta_{ij}\mu_j(t, B) + \sum_{k=1}^{n} (\mu_j * A_k)(t) + (Q * H_{ij})(t),$$

where $A_k = A_1 * A^{*(k-1)}$, and, besides,

$$0 \leqslant (\mu_j * A_k)(t) \leqslant \int_0^t [1 - B_j(t-x)]\, dA_k(x)$$

$$= \int_0^{t_0} [1 - B_j(t-x)]\, dA_k(x) + \int_{t_0}^{t} [1 - B_j(t-x)]\, dA_k(x)$$

$$\leqslant [1 - B_j(t-t_0)] + A_k(t) - A_k(t_0) \to 0$$

at $t \to \infty$, with $t = 2t_0$, whereas

$$1 - B_j(x) = P(z_{n+1} > x \mid \zeta_n = j).$$

It should also be taken into account that

$$\int_0^\infty Q(t)\, dt = \int_0^\infty \Big[\int_0^t \mu_j(t-x, B)\, dF(x) \Big] dt = \int_0^\infty \mu_j(x, B)\, dx.$$

Proof of the theorem is similar to that of Corollary of § 2, Ch. 5. 1°. Let C1 be fulfilled, in which case the function $Q = \mu_j$ satisfies the following conditions, viz.,

(1) it is Borel measurable,

(2) it is Riemann integrable in the usual sense on $[0, \infty)$, and

(3) $0 \leqslant Q(t) \leqslant G(t) = 1 - B_j(t)$, where the function G is directly Riemann integrable on $[0, \infty)$.

It follows from the same conditions that the function Q is directly Riemann integrable.

2°. Now, let C3 be fulfilled. From (2) and (3), we have

$$P(\xi(t) \in B, \zeta(t) = j \mid \zeta_0 = i) = \delta_{ij}\mu_j(t, B) + \sum_{s=1}^{n} \mu_j * F_{ij}^s(t) + \sum_{m \geqslant 1} \mu_j * \sum_k F_{ik}^n * F_{kj}^m(t)$$

$$= \delta_{ij}\mu_j(t, B) + \sum_{s=1}^{n} \mu_j * F_{ij}^s(t) + \sum_k Q_k * H_{kj}(t),$$

where

$$Q_k = \mu_j * F_{ik}^n.$$

Since

$$\mu_j(t, B) \to 0, \quad (\mu_j * F_{ij}^s)(t) \to 0$$

as $t \to \infty$, it suffices to verify that

$$\sum_k Q_k * H_{kj}(t) \to a_j \int_0^\infty \mu_j(x, B)\, dx \tag{5}$$

as $t \to \infty$. Because

$$H_{kj}(t) \leqslant 1 + H_{jj}(t), \quad F_{ik}^n \leqslant p_{ik}^n,$$

we have

$$\sum_{k>N} Q_k * H_{kj}(t) \leqslant \Big(\sum_{k>N} p_{ik}^n \Big) \cdot \mu_j * (1 + H_{jj})(t).$$

Besides,

$$\mu_j * (1 + H_{jj})(t) \leqslant (1 - B_j) * (1 + H_{jj})(t) = 1 - B_j(t) + (1 - B_j) * H_{jj}(t);$$

therefore, by the key renewal theorem (see § 5),

$$(1 - B_j) * H_{jj}(t) \to a_j \int_0^\infty [1 - B_j(x)]\, dx = a_j w_j,$$

$$\mu_j * (1 + H_{jj})(t) \to a_j w_j \text{ as } t \to \infty,$$

and

$$\sum_{k>N} Q_k * H_{kj}(t) \leqslant \Big(\sum_{k>N} p_{ik}^n \Big) \cdot (a_j w_j + \varepsilon) \tag{6}$$

when $t > t_\varepsilon$.

We now consider the growth of each addend in (5) when $t \uparrow \infty$. We shall see below that Q_k is a continuous function. Besides,

$$0 \leqslant Q_k(t) \leqslant \int_0^t [1 - B_j(t-x)]\, dF_{ik}^n(x) = G(t),$$

and the function G is directly Riemann integrable on $[0, \infty)$, since

$$G(t) = G_1(t) - G_2(t),$$

$$G_1(t) = p_{ik}^n - \int_0^t B_j(t-x)\, dF_{ik}^n(x), \quad G_2(t) = p_{ik}^n - F_{ik}^n(t),$$

and each of the functions G_1 and G_2 is monotonic and integrable. Hence, conditions (1)-(3) of the previous item are fulfilled, with the function Q_k being directly Riemann integrable on $[0, \infty)$. Therefore, it follows from the key renewal theorem (see § 5) that

$$(Q_k * H_{kj})(t) \to a_j \int_0^\infty Q_k(t)\, dt = a_j \int_0^\infty \Big[\int_0^t \mu_j(t-x, B)\, dF_{ik}^n(x) \Big] dt \tag{7}$$

$$= a_j \int_0^\infty \mu_j(x, B)\, dx \cdot F_{ik}^n(+\infty) = a_j p_{ik}^n \int_0^\infty \mu_j(x, B)\, dx$$

as $t \to \infty$, and we obtain (5) from (6) and (7).

Thus, it now suffices to verify that the function Q_k which we denote by Q is continuous. The proof of this fact coincides *verbatim* with that of the corresponding statement in § 2, Ch. 5, where we should put $F_{kj}^n = A^{(n)}$.

§ 7. Example of Application of Regenerative Processes with Dependent Regeneration Cycles of Markov Type

Theorem. For a SMP $\zeta(t)$, $t \geqslant 0$, which satisfies A1-A3, the equality holds, viz.,

$$\lim_{t\to\infty} P\{u(t) > x,\ v(t) > y,\ \zeta_{n(t)} = i,\ \zeta_{n(t)+1} = j \mid \zeta_0 = i_0\} \tag{1}$$

$$= \frac{\pi_i}{\sum\limits_k \pi_k w_k} p_{ij} \int_{x+y}^{\infty} [1 - A_{ij}(u)]\, du,$$

where

$\{p_{ij}\}$ is the transition probability matrix for the embedded Markov chain,

A_{ij} the DF of sojourn at a state i up to the transition to a state j in the case $p_{ij} > 0$; if, however, $p_{ij} = 0$, then we can assume, e.g., that $A_{ij}(u) = 1$ for $u > 0$.

w_k the average time of sojourn at a state k up to the transition to the next state, i.e.,

$$w_k = \sum_j p_{kj} \int_0^{\infty} [1 - A_{kj}(u)]\, du,$$

$\{\pi_k\}$ any non-zero non-negative solution of the system

$$\pi_j = \sum_i \pi_i p_{ij}, \quad j \in E, \tag{2}$$

$u(t)$ the time spent from the epoch of the last jump in the process up to an epoch t,

$v(t)$ the residual time from the epoch t to the epoch of the next jump in the process,

and

$n(t)$ the number of jumps in the process up to the epoch t (in particular, $\zeta_{n(t)} = \zeta(t)$).

Remark. A non-zero non-negative solution of system (2) exists and is unique up to a constant multiplier (see Remark in § 3). In case the stationary distribution exists, we may take $\{\pi_k\}$ as such. A natural example, where the stationary distribution does not exist, will be indicated at the end of the section.

Proof. Consider the regenerative process (in the sense of § 6)

$$\xi(t) = \{u(t),\ v(t),\ \zeta_{n(t)+1}\}, \quad t \geqslant 0.$$

Making use of the limit theorem given in the section for

$$B = (x, \infty) \times (y, \infty) \times \{j\},$$

we have

$$\mu_i(\tau, B) = P\{z_{n+1} > \tau,\ \tau > x,\ z_{n+1} - \tau > y,\ \zeta_{n+1} = j \mid \zeta_n = i\}$$

$$= \begin{cases} P(z_{n+1} > y + \tau,\ \zeta_{n+1} = j \mid \zeta_n = i & \text{when } \tau > x \\ 0 & \text{when } \tau \leqslant x \end{cases}$$

$$= \begin{cases} P_{ij}[1 - A_{ij}(\tau + y) & \text{when } \tau > x \\ 0 & \text{when } \tau \leqslant x \end{cases},$$

which gives

$$a_i \int_0^{\infty} \mu_i'(\tau, B)\, d\tau = a_i p_{ij} \int_x^{\infty} [1 - A_{ij}(\tau + y)]\, d\tau = a_i p_{ij} \int_{x+y}^{\infty} [1 - A_{ij}(u)]\, du,$$

$$a_i = \frac{\pi_i}{\sum\limits_k \pi_k w_k}.$$

In order to apply the theorem of § 6, it now remains to notice that C1 of this theorem is fulfilled.

Example. Consider a random walk of a particle on integral points of the straight line. The particle is at a state (point) i for a random time possessing the exponential distribution with mean w_i, after which it passes to the neighbouring state with probability 1/2. If $\zeta(t)$ is the state of the particle at an epoch t, then the process $\{\zeta(t),\ t \geqslant 0\}$ is semi-Markov. The embedded Markov chain is irreducible and recurrent, but there exists no stationary distribution. In our case,

$$p_{ij} = \begin{cases} 1/2, & |i - j| = 1 \\ 0, & |i - j| \neq 1 \end{cases}$$

and one of the non-zero, non-negative solutions of the system

$$\pi_j = \sum_{i=-\infty}^{\infty} \pi_i p_{ij} = \frac{1}{2}(\pi_{j-1} + \pi_{j+1}), \quad j = 0, \pm 1, \pm 2, \ldots$$

is $\pi_j = 1$ for all j. If the series $\sum\limits_{-\infty}^{\infty} w_k$ converges, then the right-hand side of formula (1) assumes the form

$$\frac{w_i}{\sum\limits_{-\infty}^{\infty} w_k} p_{ij} e^{-\frac{x+y}{w_i}}.$$

Problems

101. Let $F_1, F_2, \ldots$ be a sequence of DF's. Show that there exists a sequence of independent RRV's $\xi_1, \xi_2, \ldots$ such that $P(\xi_n < x) = F_n(x)$ for all $n \geqslant 1$ and $x \in R^1$.

102. Consider two numbers μ and σ. Show that there exists a sequence of independent RRV's $\xi_1, \xi_2, \ldots$ such that $E\xi_n^2 < \infty$, and $E\xi_n = \mu$, $\operatorname{var} \xi_n = \sigma^2$ for all $n \geqslant 1$.

103. Consider a matrix $P = \{p_{ij} : i, j = 0, 1, 2, \ldots\}$ such that $p_{ij} \geqslant 0 \ \forall\, i, j$, and $\sum_j p_{ij} = 1 \ \forall\, i$. Show that there exists a sequence of RV's $\xi_1, \xi_2, \ldots$ assuming values from $E = \{0, 1, 2, \ldots\}$, and forming a homogeneous Markov chain, i.e.,

$$P(\xi_{n+1} = i_{n+1} \mid \xi_1 = i_1, \ldots, \xi_n = i_n) = P(\xi_{n+1} = i_{n+1} \mid \xi_n = i_n) = p_{i_n i_{n+1}}$$

for any integer $n \geqslant 1$, any numbers $i_1, \ldots, i_{n+1}$ from E, provided $P(\xi_1 = i_1, \ldots, \xi_n = i_n) \neq 0$.

104. Consider a family of matrices $p^t = \{p_{ij}^t : i, j = 0, 1, 2, \ldots\}$, $t > 0$, such that $p_{ij}^t \geqslant 0 \ \forall\, i, j, t$, $\sum_j p_{ij}^t = 1 \ \forall\, i, t$, and $p^{s+t} = p^s \cdot p^t$ (i.e., $p_{ij}^{s+t} = \sum_k p_{ik}^s \cdot p_{kj}^t$) $\forall\, s, t > 0$. Show that there exists a random process $\{\xi(t), t > 0\}$ with values from $E = \{0, 1, 2, \ldots\}$ forming a homogeneous Markov chain (i.e., $P\{\xi(t_{n+1}) = i_{n+1} \mid \xi(t_1) = i_1, \ldots, \xi(t_n) = i_n\} = P\{\xi(t_{n+1}) = i_{n+1} \mid \xi(t_n) = i_n\}$ for any integer $n \geqslant 1$, any numbers $i_1, \ldots, i_{n+1}$ from E and any positive numbers $t_1 < \ldots < t_{n+1}$, provided $P\{\xi(t_1) = i_1, \ldots, \xi(t_n) = i_n\} \neq 0$), while $P\{\xi(s+t) = j \mid \xi(s) = i\} = p_{ij}^t \ \forall\, i, j \in E$; $s > 0$, $t > 0$.

105. Given a certain set T, vector $\mu = \{\mu_t, t \in T\}$ and symmetric, positive semi-definite matrix $A = \{a_{st} : s, t \in T\}$ (which means that, for any vector $x = \{x_t, t \in T\}$ with only a finite number of coordinates x_t different from zero, the inequality $(Ax, x) = \sum_{s,t} a_{st} x_s x_t \geqslant 0$ holds). Show that there exist a probability space (Ω, S, P) and random process $\{\xi(t), t \in T\}$ such that

(1) the finite-dimensional distributions of the process are normal, and

(2) $E\xi(t) = \mu_t$, $\operatorname{cov}(\xi(s), \xi(t)) = a_{st} \ \forall\, s, t \in T$.

Such a process is said to be Gaussian.

106. Consider a Gaussian process $\{\xi(t) : t \in T = [0, \infty)\}$ such that

$$E\xi(t) = \mu \cdot t, \quad \operatorname{cov}(\xi(s), \xi(t)) = \sigma^2 \cdot \min(s, t) \ \forall\, s, t \geqslant 0$$

(such a process is said to be Brownian, or Wiener, with shift coefficient μ and diffusion $\sigma > 0$). We show that this process has independent increments, i.e., the RV's

$$\xi(t_2) - \xi(t_1), \ldots, \quad \xi(t_n) - \xi(t_{n-1})$$

are independent for any numbers $0 \leqslant t_1 < t_2 < \ldots < t_n$.

107. Show that there exists a Gaussian process $\{\xi(t) : t \in T = (-\infty, \infty)\}$ such that

$$\operatorname{cov}(\xi(s), \xi(t)) = e^{-|t-s|}$$

(such a process is said to be Gaussian-Markov).

108. Show that each probability measure P on a metric space is regular, i.e., if B is a Borel set in this space, and $\varepsilon > 0$, then there exists a closed set F and an open set G such that $F \subset B \subset G$, and $P(G \setminus F) < \varepsilon$.

109. Show that each probability measure P on a Polish (complete, separable and metric) space is dense, i.e., for each $\varepsilon > 0$, there exists such a compact set K in this space that $P(K) > 1 - \varepsilon$.

110. Let P be a probability measure on a Polish space. Show that if B is a Borel set in this space, and $\varepsilon > 0$, then there exists a compactum K such that $K \subset B$, while $P(B \setminus K) < \varepsilon$.

111. Let a sequence $\{\xi_t : t = 0, 1, 2, \ldots\}$ of RV's form a homogeneous Markov chain. Put

E_- = the "past" time = $\{\xi_0 = i_0, \ldots, \xi_{t-1} = i_{t-1}\}$

E_0 = the "present" time = $\{\xi_t = i_t\}$

E_+ = the "future" time = $\{\xi_{t+1} = i_{t+1}, \ldots, \xi_n = i_n\}$.

Show that if $P(\xi_k = i_k \ \forall\, k < n) \neq 0$, then we have

$$P(E_- E_+ \mid E_0) = P(E_- \mid E_0) \cdot P(E_+ \mid E_0)$$

(i.e., the "past" and "future" times are independent for the fixed "present" time).

112. Let RV's $\xi_0, \xi_1, \ldots$ be independent, and each assume values ± 1 with probabilities 1/2.

(a) Does the sequence of RV's

$$\eta_n = \frac{1}{2}(\xi_n + \xi_{n+1}), \quad n = 0, 1, 2, \ldots,$$

form a Markov chain?

(b) The same question concerning the sequence

$$\zeta_n = \xi_n \cdot \xi_{n+1}, \quad n = 0, 1, 2, \ldots.$$

113. Let $\{\xi_n : n = 0, 1, 2, \ldots\}$ be a Markov chain with the state space $E = \{1, 2, 3\}$ and transition probability matrix

$$P = \begin{bmatrix} 0 & \alpha & 1-\alpha \\ \alpha & 0 & 1-\alpha \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

Defining a sequence η_n, $n = 0, 1, 2, \ldots$ by putting

$$\eta_n = \begin{cases} 1 \text{ when } \xi_n < 3 \\ 2 \text{ when } \xi_n = 3 \end{cases},$$

show that the sequence $\{\eta_n\}$ forms a Markov chain.

114 (Sufficient Condition for the Existence of the Stationary Distribution). For a homogeneous, contracting Markov chain to possess the stationary distribution, it is sufficient that there should exist a number $\varepsilon > 0$, $s \in T$, a finite set $E_0 \subset E$, and a set of non-negative numbers $(x_0, x_1, \ldots)$ such that

$$\sum_{j>0} p_{ij}^s x_j \leqslant x_j - \varepsilon \text{ for } i \notin E_0,$$

and

$$\sum_{j \geqslant 0} p_{ij}^s x_j < \infty \text{ for } i \in E_0.$$

Hint: We can assume that $E_0 = \{0, 1, \ldots, i_0\}$. Put

$$x_i^{(1)} = x_i, \quad x_i^{(n+1)} = \sum_{j \geqslant 0} p_{ij}^s x_j^{(n)}.$$

Show that

$$x_i^{(n+1)} \leqslant x_i^{(2)} + \sum_{j=0}^{i_0} (p_{ij}^{1 \cdot s} + \ldots p_{ij}^{(n-1)s})(x_j^{(2)} - x_j + \varepsilon) - (n-1)\varepsilon,$$

whence

$$0 \leqslant \frac{x_i^{(2)}}{n-1} + \sum_{j=0}^{i_0} \frac{1}{n-1} (p_{ij}^{1 \cdot s} + \ldots + p_{ij}^{(n-1)s})(x_j^{(2)} - x_j + \varepsilon) - \varepsilon.$$

115. Let $\{\nu_n : n \geqslant 1\}$ be i.i.d. RV's assuming non-negative integer values. Forming a Markov chain $\{\xi_n : n \geqslant 0\}$ by putting

$$\xi_n = (\xi_{n-1} - 1)^+ + \nu_n, \quad n \geqslant 1, \quad \xi_0 = 0,$$

show that if $E\nu_n = \rho < 1$, then the Markov chain has the stationary distribution $\{\pi_j : j \geqslant 0\}$, and

$$\lim_{n \to \infty} P(\xi_n = j) = \pi_j, \quad \sum_{j \geqslant 0} \pi_j = 1.$$

Hint: It suffices to assume $\rho > 0$. Put

$$q_k = P(\nu_n = k).$$

From $0 < E\nu_n < 1$, it follows that $0 < q_0 < 1$.

Case 1: $q_0 + q_1 = 1$. Then (with probability one)

$$\xi_n = \nu_n = 0 \text{ or } 1, \text{ and } \pi = (q_0, q_1, 0, 0, \ldots).$$

Case 2: $q_0 + q_1 < 1$. Let $q_m > 0$ for a certain $m \geqslant 2$. Then the chain is homogeneous, irreducible and contracting (take into account

that $q_0 > 0$, $q_m > 0$). Further, use the previous problem. Put $x_j = j$, $j = 0, 1, 2, \ldots$; then, if $p_{ij} = P(\xi_{n+1} = j / \xi_n = i)$, we have:

$$\sum_{j \geqslant 0} p_{ij} x_j = E(\xi_{n+1} / \xi_n = i) = \begin{cases} E\nu_{n+1} = \rho & \text{when } i = 0 \\ i - 1 + E\nu_{n+1} = i - 1 + \rho & \text{when } i > 0 \end{cases}.$$

116. Let the transition probability matrix $P = \{p_{ij}\}$ of a homogeneous discrete-time Markov chain with a finite state space $E = \{1, \ldots, m\}$ be doubly stochastic, i.e.,

$$\sum_j p_{ij} = 1 \ \forall i \text{ and } \sum_i p_{ij} = 1 \ \forall j.$$

Show that uniform (equally likely) distribution on E is stationary.

117. (*continued*). Show that the same result also holds for the case of a homogeneous continuous-time Markov chain if, for each $t > 0$, the transition probability matrix $P^t = \{p_{ij}^t\}$ is doubly stochastic.

118. Let, for a homogeneous Markov chain with a countable state space $E = \{0, 1, 2, \ldots\}$ and n-step transition probability matrix $P^n = \{p_{ij}^n\}$, there exist

$$\lim_{n \to \infty} p_{ij}^n = \pi_j \ \forall i, j,$$

not depending on i. Then

(1) $\sum_i \pi_i \leqslant 1$, $\pi_j = \sum_i \pi_i p_{ij}$, $p_{ij} = p_{ij}^1$,

(2) $\sum_i \pi_i = 0$ or 1,

and

(3) if $\pi_i = 0 \forall_i$, there exists no stationary distribution, whereas if $\sum_i \pi_i = 1$, we obtain that $\pi = (\pi_0, \pi_1, \ldots)$ is the unique stationary distribution.

119. Let, for a homogeneous Markov chain with a finite state space and n-step transition probability matrix $P^n = \{p_{ij}^n\}$, there exist

$$\lim_{n \to \infty} p_{ij}^n = \pi_j > 0 \ \forall i, j, \ \sum_j \pi_j = 1,$$

not depending on i (such a distribution $\{\pi_j\}$ is said to be **ergodic**). Show that there exist two numbers C and ρ such that $0 < \rho < 1$, and

$$| p_{ij}^n - \pi_j | \leqslant C\rho^n \ \forall \ i, j, n.$$

120. Let $\{\xi_n\}$ and $\{\bar{\xi}_n\}$ be two homogeneous Markov chains with one and the same finite state space and the same transition probability matrix $P = \{p_{ij}\}$, but, possibly, with different initial distributions (of RV's ξ_0 and $\bar{\xi}_0$). Assume that

$$\min_{i,j} p_{ij} \geqslant \Delta > 0.$$

Show that

$$\sum_i |\bar{p}_i^n - p_i^n| \leqslant 2(1-\varepsilon)^n,$$

where

$$\bar{p}_i^n = P(\bar{\xi}_n = i), \quad p_i^n = P(\xi_n = i).$$

Renewal processes

Let $\{z_k : k \geqslant 1\}$ be a sequence of non-negative and i.i.d. RV's. Put

$$t_0 = 0, \quad t_n = z_1 + \ldots + z_n, \quad n \geqslant 1,$$
$$\nu(t) = \sup\{n : t_n < t\}, \quad t \geqslant 0.$$

The process $\{\nu(t), t \geqslant 0\}$ is called a recurrent renewal process determined by a DF $A(x) = P(z_k < x)$.

The basic assumption is that $P(z_k = 0) < 1$.

Hence, for any t, there is a moment of any order of the RV $\nu(t)$. The function $H(t) = E\nu(t)$, $t \geqslant 0$, is called a renewal function, and satisfies the renewal equation

$$H(t) = A(t) + \int_0^t H(t-x)\, dA(x).$$

Note that

$$H(t) = \sum_{n \geqslant 1} A_n(t),$$

where $A_n(t) = P(t_n < t)$.

The subject matter of renewal process theory is mainly composed of the following three statements.

T1 (Elementary Renewal Theorem).

$$\frac{H(t)}{t} \to a \text{ when } t \uparrow \infty, \text{ with } a^{-1} = \int_0^\infty x\, dA(x) \leqslant \infty.$$

T2 (Blackwell Theorem). If A is a non-arithmetic distribution, then, for any number h,

$$H(t+h) - H(t) \to ah \text{ when } t \uparrow \infty.$$

T3 (Key Renewal Theorem). If Q is a function directly Riemann integrable on $[0, \infty)$, and A a non-arithmetic distribution, then

$$\int_0^t Q(t-x)\, dH(x) \to a \int_0^\infty Q(x)\, dx \text{ when } t \uparrow \infty.$$

Corollary. Under the conditions of Theorem 3, if W is a solution of the equation

$$W = Q + A * W,$$

$W(x) = 0$ for $x < 0$, where $A * W$ is the Lebesque-Stieltjes convolution of the functions A and W, i.e.,

$$(A * W)(t) = \int_0^t A(t-x)\, dW(x) = \int_0^t W(t-x)\, dA(x),$$

then

$$\lim_{t \to \infty} W(t) = a \int_0^\infty Q(x)\, dx.$$

121. Show that

$$\{\nu(t) \geqslant n\} = \{z_1 + \ldots + z_n < t\}, \quad n \geqslant 1,$$

and

$$P\{\nu(t) = n\} = P_n(t) = A_n(t) - A_{n+1}(t), \quad n \geqslant 0.$$

122. Prove that the inequality

$$P(z_k = 0) < 1$$

is equivalent to

$$P\{\nu(t) < \infty\} = 1 \ \forall t \geqslant 0.$$

123. Let $A(x) = 1 - e^{-ax}$, $x \geqslant 0$, $a > 0$, in which case the renewal process $\{\nu(t), t \geqslant 0\}$ is said to be Poisson (with intensity a). Show that

$$P_n(t) = \frac{(at)^n}{n!} e^{-at}, \quad t > 0, \quad n = 0, 1, 2, \ldots .$$

124. Let

$$A(x) = \int_0^{cx} \frac{u^{r-1}}{(r-1)!} e^{-u}\, du, \quad x \geqslant 0, \quad c > 0,$$

in which case the renewal process is said to be Erlang of order $r \geqslant 1$ (r is an integer). Show that

$$P_n(t) = \sum_{k=nr}^{nr+r-1} \frac{(ct)^k}{k!} e^{-ct}.$$

Hint: An RV having the Erlang distribution of order r is representable as the sum of RV's i.i.d. according to the exponential law.

125. Let, for a number $c > 0$,

$$P(z_k = c) = p, \text{ and } P(z_k = 0) = q = 1 - p.$$

Find the generating function of the RV $\nu(t)$, and renewal function $H(t)$.

126. Consider a sequence $\{z_k : k \geqslant 1\}$ of positive semi-definite RV's, and put

$$t_0 = 0, \ t_n = z_1 + \ldots + z_n, \ n \geqslant 1,$$
$$\nu(t) = \sup\{n : t_n < t\}, \ t \geqslant 0.$$

Prove that the relations

$$\frac{1}{n}\sum_{1}^{n} z_k \xrightarrow{P} 1/a \text{ as } n\to\infty$$

and

$$\frac{\nu(t)}{t} \xrightarrow{P} a \text{ as } t\to\infty$$

are equivalent.

127. Let A be a non-arithmetic distribution, and

$$\mu_2 = \int_0^\infty x^2\, dA(x) < \infty.$$

Show that

$$H(t) - at \to \frac{a^2\mu_2}{2} - 1 \text{ as } t\to\infty.$$

Hint: Putting

$$W(t) = H(t) + 1 - at \text{ for } t\geqslant 0, \text{ and } Q(t) = a\int_t^\infty [1 - A(u)]\, du,$$

we have:

$$W = Q + A * W.$$

Besides,

$$a^{-1} = \int_0^\infty x\, dA(x) = \int_0^\infty [1 - A(x)]\, dx, \quad \mu_2 = \int_0^\infty [1 - A(x)]\, dx^2.$$

128. Starting with the moment $t = 0$, buses arrive at a bus stop with time intervals $z_1, z_2, \ldots$. Assume that the RV's $z_1, z_2, \ldots$ are i.i.d., and

$$A(x) = P(z_k < x),\ x\geqslant 0;\ P(z_k = 0) < 1,$$

$$a^{-1} = \int_0^\infty x\, dA(x) < \infty.$$

Denote the bus waiting time by $\xi(t)$, starting with an epoch $t \geqslant 0$. Show that

(1) if A is the arithmetic distribution, then, for a certain $x > 0$, there exists no

$$\lim_{t\to\infty} P\{\xi(t) < x\},$$

and

(2) if A is a non-arithmetic distribution, then there exists

$$\lim_{t\to\infty} P\{\xi(t) < x\} = a\int_0^x [1 - A(u)]\, du.$$

Hint: By the total probability formula,

$$P\{\xi(t) < x\} = \sum_{n\geqslant 0} P\{t_n \leqslant t < t_{n+1},\ \xi(t) < x\},$$

whence, taking into account that

$$\xi(t) = t_{n+1} - t = z_{n+1} - (t - t_n) \text{ for } t_n \leqslant t < t_{n+1},$$

and

$$P\{t_n \leqslant t < t_{n+1},\ \xi(t) < x \mid t_n = u\} = P\{u \leqslant t < u + z_{n+1}, \text{ and } z_{n+1} - (t-u) < x\} = P\{(t-u) < z_{n+1} < x + (t-u),\ t \geqslant u\} = Q(t-u)$$

$$= \begin{cases} A(x+t-u) - A(t-u) & \text{for } t-u \geqslant 0 \\ 0 & \text{for } t-u < 0 \end{cases},$$

we have:

$$P\{\xi(t) < x\} = P\{0 \leqslant t < t_1,\ \xi(t) < x\} + \sum_{n\geqslant 1}\int_0^t Q(t-u)\, dA_n(u)$$

$$= g(t) + \int_0^t Q(t-u)\, dH(u),$$

where

$$0 \leqslant g(t) = P\{0 \leqslant t < t_1,\ \xi(t) < x\} \leqslant P\{t < z_1\} \to 0 \text{ as } t\to\infty.$$

129. Let Q be a Borel measurable and Lebesgue integrable function on $[0, \infty)$. Show that

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \left\{\int_0^t Q(t-u)\, dH(u)\right\} dt = a\int_0^\infty Q(t)\, dt.$$

Hint: Changing the order of integration, verify that

$$\frac{1}{T}\int_0^T\int_0^t Q(t-u)\, dH(u)\, dt = \frac{H(T)}{T}\int_0^T Q(v)\, dv + \varepsilon_T,$$

and

$$|\varepsilon_T| \leqslant \frac{H(T)}{T}\int_{T-u_0}^\infty |Q(v)|\, dv + \frac{H(u_0)}{T}\int_0^\infty |Q(v)|\, dv, \quad 0 \leqslant u_0 \leqslant T.$$

130. Prove that, for $h \geqslant 0$,

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T [H(t+h) - H(t)]\, dt = ah.$$

Regenerative processes

If $\xi = \{\xi(t) : 0 \leqslant t < z\}$ is a random process on an interval $[0, z)$ of random length z, then the pair (z, ξ) is called a **cycle.** Let $\{(z_k, \xi_k) : k \geqslant 1\}$ be a sequence of independent, stochastically equivalent cycles. In particular, this means that the sequence $\{z_k : k \geqslant 1\}$ forms the renewal process

$$\nu(t) = \sup\{n : t_n < t\}, \quad t \geqslant 0,$$

$$t_0 = 0, \; t_n = z_1 + \ldots + z_n, \quad n \geqslant 1.$$

Proceeding from such cycles (z_k, ξ_k), $k \geqslant 1$, we form a random process $\xi(t)$, $t \geqslant 0$, by putting

$$\xi(t) = \xi_k(t - t_{k-1}) \text{ if } t_{k-1} \leqslant t < t_k$$

(see Fig. 10).

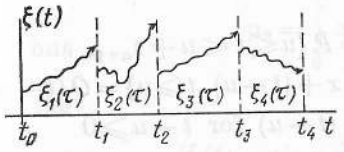

Fig. 10

Such a process $\{\xi(t), t \geqslant 0\}$ is said to be regenerative, whereas $\{t_k\}$ are termed **regeneration points.**

Let B be a certain measurable state space of a process $\xi(t)$, $t \geqslant 0$. The goal of the theory of regenerative processes is to indicate the conditions for the existence of

$$(*) \qquad \lim_{t\to\infty} P\{\xi(t) \in B\} = P_B,$$

and also a method of finding it. In this connection, we put

$$\mu_B(\tau) = P\{\xi_1(\tau) \in B, \; z_1 > \tau\}, \; \tau \geqslant 0.$$

Theorem. Assume that a DF $A(x) = P(z_1 < x)$, $x \geqslant 0$, is non-arithmetic, and $a^{-1} = \int_0^\infty x dA(x) = Ez_1 < \infty$, whereas a function $\tau \to \mu_B(\tau)$ is Riemann integrable on each finite interval. Then there exists the limit (*), and

$$P_B = a\int_0^\infty \mu_B(\tau)\, d\tau.$$

131. Let the intervals of operation of a certain device alternate with those of repair. Denote by $\{u_k, k \geqslant 1\}$ the sequence of the intervals during which the device is in service, and by $\{v_k\}$ the corresponding sequence of intervals during which it is out of service. Assume that the RV's $\{u_k\}$ and $\{v_k\}$ are i.i.d. inside each sequence. Let $\xi(t) = 0$ if the device is being repaired at an epoch t, and $\xi(t) = 1$ otherwise. Assuming that $Eu_1 + Ev_1 < \infty$, and the distribution $A(x) = P(u_1 + v_1 < x)$ non-arithmetic, show that

$$\lim_{t\to\infty} P\{\xi(t) = 1\} = \frac{Eu_1}{Eu_1 + Ev_1}.$$

Hint: $\xi(t)$ is a regenerative process with regeneration epochs t_1, t_2, ..., where

$$t_k = z_1 + \ldots + z_k, \; z_k = u_k + v_k,$$

$$B = \{1\}, \; \mu_B(t) = P\{\xi(t) = 1, \; z_1 > t\}$$

$$= P(u_1 > t, \; u_1 + v_1 > t) = P(u_1 > t),$$

and

$$\int_0^\infty \mu_B(t)\, dt = \int_0^\infty P(u_1 > t)\, dt = Eu_1.$$

132. Starting with the epoch $t = 0$, buses arrive at a stop with time intervals $z_1, z_2, \ldots$. We assume that the RV's $z_1, z_2, \ldots$ are i.i.d., whereas

$$A(x) = P(z_k < x), \; x \geqslant 0,$$

and

$$0 < a^{-1} = \int_0^\infty x\, dA(x) < \infty.$$

Let $\xi(t)$ be the next bus waiting time, starting with an epoch t, and $\eta(t)$ the time passed since the arrival of the last bus up to the epoch t. Note that

$$\eta(t) = t - t_{\nu(t)}, \; \xi(t) = t_{\nu(t)+1} - t.$$

Show that if A is a non-arithmetic distribution, then

$$\lim_{t\to\infty} P\{\xi(t) \geqslant x, \; \eta(t) \geqslant y\} = a\int_{x+y}^\infty [1 - A(u)]\, du.$$

Hint: In our case, when $B = [x, \infty) \times [y, \infty)$,

$$\mu_B(t) = P\{\xi(t) \geqslant x, \; \eta(t) \geqslant y, \; z_1 > t\}$$

$$= P(z_1 - t \geqslant x, \; t \geqslant y, \; z_1 > t)$$

$$= P(z_1 \geqslant x + t, \; t \geqslant y)$$

$$= \begin{cases} 1 - A(x+t) & \text{for } t \geqslant y \\ 0 & \text{for } t < y \end{cases}.$$

133 (*continued*). If A is a non-arithmetic distribution, then, as $t \to \infty$,

$$(1) \qquad \lim P\{\xi(t) + \eta(t) < x\} = A^*(x) = a\int_0^x [A(x) - A(t)]\, dt,$$

$$(2) \qquad \lim P\{|\xi(t) - \eta(t)| < x\} = a\int_0^x [1 - A(u)]\, du,$$

and

$$\lim P\left\{\frac{\eta(t)}{\xi(t)+\eta(t)}<x\right\}=x \text{ for } 0\leqslant x\leqslant 1. \tag{3}$$

134 (*continued*). Show that

$$\lim_{t\to\infty} E\zeta(t)=\int_0^\infty x dA^*(x)=a\mu_2, \quad \mu_2=\int_0^\infty x^2 dA(x),$$

where $\zeta(t)=\xi(t)+\eta(t)$.

RV convergence

135. Prove that a.c. convergence $\xi_n\to\xi$ is equivalent to $\eta_n=\sup_{m\geqslant n}|\xi_m-\xi|\to 0$ in probability.

136. Prove that for a sequence of RV's $\{\xi_n\}$ to be convergent a.c., it is sufficient that

$$\sum_{n\geqslant 1} P\{|\xi_{n+1}-\xi_n|>\varepsilon_n\}<\infty$$

for a certain summable sequence of non-negative numbers $\{\varepsilon_n: n\geqslant 1\}$.

137. Prove that, for a discrete probability space, convergence a.c. coincides with that in probability.

138. Prove that, for a discrete probability space, convergence in the mean implies convergence a.c.

139. Show that convergence $\xi_n\downarrow 0$ in probability implies convergence a.c. $\xi_n\to 0$.

140. Let $\{F_n,\ n\geqslant 0\}$ be a sequence of continuous DF's and $F_n\to F_0$ weakly. Show that there exist such a probability space (Ω, S, P) and sequence $\{\xi_n,\ n\geqslant 0\}$ of RV's on it that

(1) F_n is the DF of the RV ξ_n for all $n\geqslant 0$, and

(2) $\xi_n\to\xi_0$ a.c.

141. Show that $\xi_n\to c=\text{const}$ in probability is equivalent to $\xi_n\to c$ in distribution.

142. Let f be a continuous function on R^s, while, e.g., $s=1$. Show that

(1) $\xi_n\to\xi_0$ in probability implies $f(\xi_n)\to f(\xi_0)$ in probability, and

(2) $\xi_n\to\xi_0$ in distribution implies $f(\xi_n)\to f(\xi_0)$ in distribution (ξ_n being random vectors of dimension s).

143. Consider the following algorithm of the solution to a mathematical programming problem

$$\inf_{x\in X} f(x)=c,$$

where $(X, \mathcal{A})$ is a measurable space, and f a Borel function. Let $\xi_1, \xi_2, \ldots$ be a sequence of i.i.d. RV's with values in $(X, \mathcal{A})$, and,

for any $\varepsilon>0$,

$$P(f(\xi_1)\leqslant c+\varepsilon)>0.$$

Put $y_1=f(\xi_1)$, $y_{n+1}=\min(y_n, f(\xi_{n+1}))$. Show that $P(y_n\downarrow c)=1$ (e.g., if X is a bounded and measurable subset of a finite-dimensional Euclidean space, while $\xi_1, \xi_2, \ldots$ independent RV's uniformly distributed in X, i.e., $P(\xi_1\in A)=v(A)/v(X)$, where $v(A)$ is the measure of a set A, and $v\{x: f(x)\leqslant c+\varepsilon\}>0\ \forall\varepsilon>0$).

Solution. The sequence $\{y_n\}$ being monotonic, by the results of Problem 139, it suffices to verify that $y_n\downarrow c$ in probability. For any $\varepsilon>0$, we have

$$\begin{aligned} P(|y_n-c|>\varepsilon) &= P(y_n>c+\varepsilon)\\ &= [P(f(\xi_k)>c+\varepsilon,\ k=1,\ldots,n)\\ &= [P(f(\xi_1)>c+\varepsilon)]^n\to 0 \end{aligned}$$

as $n\to\infty$, because $P(f(\xi_1)>c+\varepsilon)<1$ from the data.

144. Prove that

(a) $\xi_n\to\xi$ in probability is equivalent to any subsequence $\{n'\}\subset\{n\}$ containing a subsequence $\{n''\}\subset\{n'\}$ such that $\xi_{n''}\to\xi$ a.c., and

(b) for a metric space, $x_n\to x$ is equivalent to any subsequence $\{n'\}\subset\{n\}$ containing a subsequence $\{n''\}\subset\{n'\}$ such that $x_{n''}\to x$.

It follows from these arguments that the space of RRV's (given on one and the same probability space) or that of classes of equivalent RV's with convergence a.c. is not metrizable; moreover, it is not a Hausdorff topological space for which statement (b) is also valid).

145. Let RV's $\xi_1, \xi_2, \ldots$ be i.i.d., $E\xi_1=\mu$, $\operatorname{var}\xi_1=\sigma^2<\infty$. Put

$$S_n=\xi_1+\ldots+\xi_n, \quad \zeta_n=\frac{S_n-ES_n}{\sqrt{\operatorname{var}S_n}}=\frac{1}{\sqrt{n}}\sum_1^n\frac{\xi_k-\mu}{\sigma}, \quad \sigma>0.$$

Show that there exists no such RV ζ with the normal distribution $N(0, 1)$ that $\zeta_n\to\zeta$ a.c., either in the mean, or in probability.

Hint: 1°. It suffices to verify the statement for (weakest) convergence in probability.

2°. Convergence $\eta_n\xrightarrow{\text{s.m.}}\eta$ is equivalent to $\eta_n\to\eta$ in probability, and $E\eta_n^{2i}\to E\eta^2$.

3°. Let there exist an RV ζ with the normal distribution $N(0, 1)$, and $\zeta_n\to\zeta$ in probability. Since, in addition, $1=E\zeta_n^2\to E\zeta^2=1$, according to 2°, we have $\zeta_n\xrightarrow{\text{s.m.}}\zeta$. In particular, for any $\varepsilon>0$, there exists $N=N_\varepsilon$ such that, when $m>n\geqslant N$,

$$\operatorname{var}(\zeta_m-\zeta_n)<\varepsilon.$$

However, assuming $\mu = 0$, $\sigma = 1$, we have

$$\zeta_m - \zeta_n = \left(\frac{1}{\sqrt{m}} - \frac{1}{\sqrt{n}}\right)(\xi_1 + \ldots + \xi_n) + \frac{1}{\sqrt{m}}(\xi_{n+1} + \ldots + \xi_m),$$

and, since the RV's to be added are independent,

$$\operatorname{var}(\zeta_m - \zeta_n) = 2 - 2\sqrt{\frac{n}{m}}.$$

146. Let $\xi(t)$, $t \geqslant 0$, be a Poisson process. Show that, in the sense of convergence in probability,

(1) the process $\xi(t)$, $t \geqslant 0$, is continuous, while

(2) there exists the derivative $\xi'(t)$, and $\xi'(t) = 0$ a.c. $\forall\, t$.

147. Let $\omega(t)$, $t \geqslant 0$ be a Wiener process. Show that, provided

$$\omega(t_1) = \omega_1, \quad \omega(t_2) = \omega_2, \quad t_1 < t < t_2,$$

the RV $\omega(t)$ has the normal distribution with mean

$$\omega_1 + \frac{\omega_2 - \omega_1}{t_2 - t_1}(t - t_1) \quad \text{and variance} \quad \frac{(t_2 - t)(t - t_1)}{t_2 - t_1}.$$

148. Let $\omega(t)$, $t \geqslant 0$, be a standard Wiener process, and g a bounded, measurable function on the straight line. Show that

$$E\int_0^c g(\omega(t))\,dt = \int_0^c \left[\frac{1}{\sqrt{2\pi t}}\int_{-\infty}^{\infty} g(y)\,e^{-\frac{y^2}{2t}}\,dy\right]dt = \int_0^c Eg(\omega(t))\,dt.$$

149. Let $\omega(t)$, $t \geqslant 0$, be a standard Wiener process, and $g_1, \ldots, g_m$ functions integrable on $[0, c]$. Put $\xi = (\xi_1, \ldots, \xi_m)$, where

$$\xi_k = \int_0^c g_k(t)\,\omega(t)\,dt.$$

Show that the random vector ξ has the normal distribution $N(0, A)$, where $A = \{a_{ij}\}$, and

$$a_{ij} = \int_0^c \int_0^c g_i(s) \cdot g_j(t) \min(s, t)\,ds\,dt, \quad i, j = 1, \ldots, m.$$

150. Let $\omega(t)$, $t \geqslant 0$, be a standard Wiener process, and c a certain positive number. Show that

$$E\left|\sum_{k=0}^{n-1}[\omega(kh + h) - \omega(kh)]^2 - c\right|^2 \to 0$$

as $n \to \infty$, where $h = c/n$.

Part 3

Statistics

Introduction

§ 1. Posing Statistical Problem

Any statistical investigation deals with a collection of the results of observations x representing the values of an RV ξ with a distribution which depends on an unobservable parameter $\theta \in \Theta$. Statistical inferences form a basis to make some or other statement regarding the true value of the unobservable parameter, proceeding from the results of the observations x.

In other words, we seek θ in relation to x. The dependence between θ and x consists in x being the value of the RV ξ with a distribution P_θ which is a function of $\theta \in \Theta$.

Example. Let $x = (x_1, \ldots, x_n)$ be the value of a random vector $\xi = (\xi_1, \ldots, \xi_n)$ whose components are i.i.d. according to $N(\mu, \sigma^2)$, μ and/or σ^2 possibly being unknown. E.g., if μ and σ^2 are unknown, then $\theta = (\mu, \sigma^2)$ is taken as the unobservable parameter, and $\Theta = R^1 \times R^1_+$. This example may serve as a model of independent measurements of a certain variable whose value is unknown, and with measuring errors distributed normally.

We should distinguish between ξ and x, i.e., an RV and its value; however, in order not to introduce extra notation, we will use (where it is not a reason for ambiguity) one and the same symbol both for an RV and its value.

A stochastic automaton transforming an input into an output signal may serve as a mathematical model of a statistical experiment. To describe such an automaton, we will interpret $\Theta = \{\theta\}$ as the set of input, and $X = \{x\}$ as that of output signals.

The automaton works in accordance with the following rule, viz., if an input signal is θ, then the output will be x with a distribution P_θ (see Fig. 11). It is required to make a statement concerning the input signal θ on the basis of the output signal x and the "transfer function" P_θ.

The set $X = \{x\}$ of possible observations is called a **sample space.** Actually, it is a measurable space $(X, \mathcal{A})$, so that, for each set $A \in \mathcal{A}$, we may speak of the probability $P_\theta(x \in A)$ that an observation x is contained in this set A, whereas the set $\Theta = \{\theta\}$ is called the **parameter space.**

Denote by $D = \{d\}$ the set of possible decisions relative to the true value of the parameter θ (decision space). Any statistical analysis should make it possible to make a decision $\delta(x) \in D$ (regarding the true value of parameter θ) on the basis of an observation x. The function δ: $X \to D$ is called a **decision function,** or **decision rule** (DR), or **(statistical) strategy.**

It is natural that the final result of any statistical analysis should be a strategy δ, viz., if x is an observation, then $\delta(x)$ is a decision (regarding the true value of the parameter θ).

As a rule, we first confine ourselves to a certain set $\Delta = \{\delta\}$ of admissible strategies, and do not consider all the strategies δ: $X \to D$.

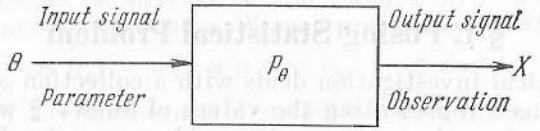

Fig. 11

Case 1. A decision may be a point from Θ, and $D = \Theta$. The statistical analysis of this case is **point estimation theory.** E.g., if $x = (x_1, \ldots, x_n)$, and the RV's $x_1, \ldots, x_n$ are i.i.d. according to $N(\theta, 1)$, then the strategy

$$x \to \delta(x) = \frac{1}{n}(x_1 + \ldots + x_n) = \bar{x}$$

called the **estimator** of parameter θ possesses the following properties, viz.,

(1) $E_\theta \delta(x) = \theta \; \forall \; \theta$, and

(2) $E_\theta \mid \delta(x) - \theta \mid^\alpha \leqslant E_\theta \mid \delta_1(x) - \theta \mid^\alpha$ for all θ, $\alpha \geqslant 1$, and all the strategies δ_1 satisfying property (1).

Case 2. A decision may be a certain subset of the set Θ, and $D \subset \mathcal{P}(\Theta)$, where $\mathcal{P}(\Theta)$ is the set of all subsets of the set Θ. The statistical analysis of this case is **confidence set (interval estimation) theory.** E.g., if $x = (x_1, \ldots, x_n)$, and the RV's $(x_1, \ldots, x_n)$ are i.i.d. according to $N(\theta, \sigma^2)$ with known variance σ^2, $\sigma > 0$, then the RV $\sqrt{n}\frac{\bar{x}-\theta}{\sigma}$, where $\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n)$, has the normal distribution $N(0, 1)$; therefore,

$$P_\theta\left\{|\theta - \bar{x}| \leqslant \frac{3\sigma}{\sqrt{n}}\right\} = P_\theta\left\{\left|\sqrt{n}\frac{\bar{x}-\theta}{\sigma}\right| \leqslant 3\right\}$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-3}^{3} e^{-\frac{u^2}{2}}\,du = 0.998\ldots.$$

Now, if

$$\delta(x) = \left[\bar{x} - \frac{3\sigma}{\sqrt{n}},\ \bar{x} + \frac{3\sigma}{\sqrt{n}}\right],$$

then

$$P_\theta\{\theta \in \delta(x)\} = 0.998\ldots \; \forall\, \theta.$$

Case 3. Let $\Theta = \Theta_1 + \ldots + \Theta_s$ and $D = \{d_1, \ldots, d_s\}$, where d_i is a decision of the form $\theta \in \Theta_i$. The statistical analysis of this case is **(composite) hypothesis testing theory.** Particularly interesting is a special case where the sets $\Theta_1, \ldots, \Theta_s$ are one-point, i.e., $\Theta = \{\theta_1, \ldots, \theta_s\}$, and d_i is a decision of the form $\theta = \theta_i$. In this case, the statistical analysis is **(simple) hypothesis testing theory.** E.g., these cases may serve as a mathematical idealization (model) of classification, discrimination of objects, which are completely characterized by the parameter θ, on the basis of the result of observation (measurement) of this parameter with random errors.

§ 2. Risk Function

Since the ultimate goal of any statistical investigation is the choice of a strategy, additional statements permitting us to order the strategies are necessary; in particular, to indicate, for any pair of strategies δ_1 and δ_2, which of them is preferable over the other.

One of the possibilities permitting us to order strategies partially consists in the following. Let $l(d \mid \theta)$ be losses in using a decision d if the true value of the unobservable parameter is θ. This means that a function $l : D \times \Theta \to R^1$ (l = loss) is given. Now, if a strategy $\delta \in \Delta$ is used, then the average losses are

$$R(\delta \mid \theta) = E_\theta l(\delta(x) \mid \theta) = \int_X l(\delta(x) \mid \theta)\, P_\theta(dx).$$

The function $l(\delta, \theta) \to R(\delta \mid \theta)$ is called a **risk function**, where $R(\delta \mid \theta)$ are the average losses in using a strategy δ if θ is the true value of the unobservable parameter.

Now, if, for two strategies δ_1 and δ_2,

(1) $R(\delta_1 \mid \theta) \leqslant R(\delta_2 \mid \theta)\ \forall \theta$,

then it is natural to regard the strategy δ_1 as preferable over δ_2, which is what defines partial order in the set Δ of admissible strategies.

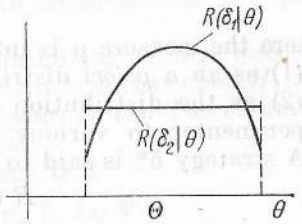

Fig. 12

Remark 1. For any $\theta \in \Theta$, the function $x \to l(\delta(x) \mid \theta)$ should be measurable with respect to measure P_θ (this is required, at the minimum, in defining a risk function). These (and other) arguments determine the choice of the set Δ of admissible strategies.

Remark 2. The statement "θ is the true value of an unobservable parameter" means that the RV (observation) x has distribution P_θ.

Unfortunately, not all pairs of admissible strategies δ_1 and δ_2 are comparable in the sense of (1) (see Fig. 12), which is the principal difficulty in posing a statistical problem.

It may happen, however, that there exists a strategy $\delta^* \in \Delta$ such that

$$R(\delta^* \mid \theta) \leqslant R(\delta \mid \theta) \ \forall\, \theta \in \Theta,\ \delta \in \Delta,$$

in which case δ^* is called an **(optimal) strategy with uniformly minimum risk**, or simply an **optimal strategy.**

§ 3. Strategy Ordering

1°. **Minimax approach.** A simple method permitting us to order a set Δ of admissible strategies consists in the following. For $\delta \in \Delta$, put

$$R(\delta) = \sup_{\theta} R(\delta \mid \theta).$$

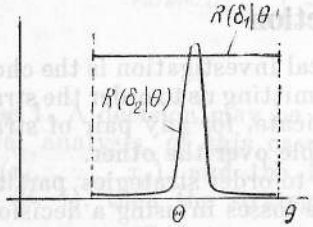

Fig. 13

We shall now regard the strategy δ_1 as preferable over δ_2 if $R(\delta_1) \leqslant R(\delta_2)$. However, if, for a strategy $\delta^* \in \Delta$, $R(\delta^*) \leqslant R(\delta)$ for all $\delta \in \Delta$, then δ^* is called a **minimax strategy**. Such a strategy is the best guarantee against great losses. Fig. 13 clarifies one defect of a minimax approach, viz., δ_1 is better than δ_2, since $\sup_{\theta} R(\delta_1 \mid \theta) < \sup_{\theta} R(\delta_2 \mid \theta)$, though the strategy δ_2 looks preferable over δ_1.

2°. **Bayes approach.** Another very important method permitting us to order strategies is as follows. For $\delta \in \Delta$, put

$$R(\delta) = \int_{\Theta} R(\delta \mid \theta)\, \mu(d\theta),$$

where the measure μ is interpreted either

(1) as an *a priori* distribution of the parameter θ or

(2) as the distribution of degree of significance ascribed by the experimenter to various θ.

A strategy δ^* is said to be **Bayes** if

$$R(\delta^*) = \min_{\delta} R(\delta).$$

There are a number of weighty arguments for such an approach. One of them consists in existing, for any strategy $\delta \in \Delta$ (under sufficiently general conditions), of a Bayes strategy δ' (associated with a certain distribution μ on Θ)preferable over δ in the sense that

$$R(\delta' \mid \theta) \leqslant R(\delta \mid \theta)\ \forall\, \theta.$$

The Bayes approach has one defect, viz. the necessity to know the *a priori* distribution of parameter θ in addition to indicating that

it is an RV (which itself requires fervid imagination in certain cases). As a rule, however, we do not possess reliable information regarding the *a priori* distribution of θ, and any interpretation of how much confidence or uncertainty we can have in a given value of θ is not amenable to a numerical estimation.

Example (*Bayes solution of the classification problem*).

A. Posing the problem. Let certain objects be grouped into s classes $H_1, \ldots, H_s$, and x a characteristic of an object under observation. In other words, we assume that x is an RV with one of s known distributions $P_1, \ldots, P_s$ according as to which of the s classes $H_1, \ldots, H_s$ the object belongs.

Denote by w_{ij} the losses due to placing an object from a class H_j in a class H_i. Finally, let π_j represent the (*a priori*) distribution that the object under observation is in a class j.

In the given situation, the decision rule is a (measurable) mapping δ: $X \to D = \{1, \ldots, s\}$, where X is the set of possible values of x (sample space). Viz., if $\delta(x) = i$, then an object with characteristic x is declared to belong to the class H_i. The problem is to select a DR δ minimizing average losses (Bayes risk). We have

$$R(\delta) = \sum_{i,\,j} \pi_j \alpha_{ij} w_{ij},$$

where $\alpha_{ij} = P_j\{\delta(x) = i\}$ is the probability that an object from a class H_j will be placed in a class H_i.

B. Basic result. Let $p_k(x)$ be the density of a distribution P_k (e.g., with respect to measure $\lambda = P_1 + \ldots + P_s$). Then

$$L_i = L_i(x) = \sum_{k=1}^{s} \frac{\pi_k p_k(x)}{\sum_{j=1}^{s} \pi_j p_j(x)}\, w_{ik} = \frac{\sum_{k=1}^{s} \pi_k p_k(x)\, w_{ik}}{\sum_{k=1}^{s} \pi_k p_k(x)}$$

are (*a posteriori*) losses if an object with characteristic x is placed in the class H_i.

It is natural to expect that the DR δ minimizing average losses consists in placing an object with characteristic x in that class H_i which is associated with minimal *a posteriori* losses $L_i(x)$. The assumption is justified by the theorem proved below.

Denote by Δ^* the class of DR's δ^* such that

$$\delta^*(x) = i \text{ implies } L_i(x) \leqslant L_j(x)\ \forall\, j.$$

Theorem. If $\delta^* \in \Delta^*$, then $R(\delta^*) \leqslant R(\delta)$ for all DR's δ.

Proof. (a) We show that, for any DR δ,

$$EL_{\delta} = EL_{\delta(x)}(x) = R(\delta),$$

where $E = \sum_{k=1}^{s} \pi_k E_k$, and E_k is the ME operator of the RV depending on the RV x with distribution P_k.

In fact, put $X_i = \{x : \delta(x) = i\}$. Then

$$\alpha_{ij} = \int 1_{X_i} p_j \, d\lambda,$$

and

$$EL = \sum_1^s \pi_j E_j \left(\sum_1^s 1_{X_i} \right) L_\delta = \sum_{i,j} \pi_j E_j 1_{X_i} L_i$$
$$= \sum_{i;j} \pi_j \int 1_{X_i} p_j L_i \, d\lambda = \sum_i \int 1_{X_i} L_i \sum_j \pi_j p_j \, d\lambda$$
$$= \sum_i \int 1_{X_i} \sum_j \pi_j p_j w_{ij} \, d\lambda = \sum_{i,j} \pi_j w_{ij} \int 1_{X_i} p_j \, d\lambda$$
$$= \sum_{i,j} \pi_j w_{ij} \alpha_{ij} = R(\delta).$$

(b) Put

$$L = L(x) = \min_i L_i(x).$$

For any $\delta^* \in \Delta$, we have $L_{\delta^*} = L$; therefore,

$$EL = EL_{\delta^*}.$$

(c) Now, for any DR δ,

$$R(\delta) = EL_\delta \geqslant E \min_i L_i = EL = EL_{\delta^*} = R(\delta^*).$$

Chapter 9

Mathematical Statistics: Basic Concepts

§ 1. Statistical Structure

Let $\mathcal{P} = \{P\}$ be a certain family of probability measures (distributions) on a measurable space $(X, \mathcal{A})$. The triplet $(X, \mathcal{A}, \mathcal{P})$ is called a **statistical structure.**

Remark 1. If $\mathcal{P}$ consists of one element, then the statistical structure turns into a probability space.

Remark 2. The family $\mathcal{P}$ is often parametrized and written in the form $\mathcal{P} = \{P_\theta, \ \theta \in \Theta\}$, whereas the set Θ is called a parameter space. The assumption $\theta_1 \neq \theta_2 \Rightarrow P_{\theta 1} \neq P_{\theta 2}$ is regarded as natural.

Remark 3. In mathematical statistics, the space $(X, \mathcal{A})$ is interpreted as the space of possible observations, or sample space. It is

then required to make a statement concerning a distribution P from $\mathcal{P}$ on the basis of an observation $x \in X$, the RV with realization x obeying P.

We will say that a statistical structure $(X, \mathcal{A}, \mathcal{P})$ is **dominated by a σ-finite measure** μ on $(X, \mathcal{A})$ if all measures $P_\theta \in \mathcal{P}$ are absolutely continuous with respect to μ, i.e., for any measure $P_\theta \in \mathcal{P}$, there exists such a function (viz., density of P_θ with respect to μ) $L(x \mid \theta)$ of $x \in X$ that

$$P_\theta(A) = \int_A L(x|\theta) \, \mu(dx) \quad \text{for all } A \in \mathcal{A},$$

in which case $X \times \Theta \ni (x, \theta) \to L(x \mid \theta) \in [0, \infty)$ is called a **likelihood function** (LF).

§ 2. Statistic

Any measurable function of observation is called a **statistic.** More precisely, if $(X, \mathcal{A}, \mathcal{P})$ is a statistical structure, then a measurable mapping T of the measurable space $(X, \mathcal{A})$ into a certain measurable space $(Y, \mathcal{B})$ is called a statistic.

Thus, for any $P \in \mathcal{P}$, a statistic T is an RV in the probability space $(X, \mathcal{A}, P)$.

Two statistics $T_1 : X \to Y$ and $T_2 : X \to Y$ are said to be **equivalent** if the event $\{T_1 \neq T_2\} = \{x : T_1(x) \neq T_2(x)\} \ (\in \mathcal{A})$ is $\mathcal{P}$-negligible. Meanwhile, an event $A \in \mathcal{A}$ is said to be **$\mathcal{P}$-negligible** if $P(A) = 0$ for all $P \in \mathcal{P}$.

Two statistics T_1 and T_2 on $(X, \mathcal{A}, \mathcal{P})$ are said to be **independent** if, for any $P \in \mathcal{P}$, the RV's T_1 and T_2 are independent when considered in the probability space (X, A, P).

A statistic T on $(X, \mathcal{A}, \mathcal{P} = \{P_\theta, \ \theta \in \Theta\})$ is said to be **integrable** if, for any distribution P_θ, $\theta \in \Theta$, the RV T considered in the probability space $(X, \mathcal{A}, P_\theta)$ is integrable. The ME of the RV T, which is associated with a distribution P_θ is denoted by $E_\theta T$, or $E_\theta T(x)$.

§ 3. Sufficient Statistic

The concept of sufficient statistic is related to the idea that the amount of information regarding the true value of a parameter θ, which is contained in an observation x, coincides with the amount of information regarding θ, and contained in $T(x)$ (expression "θ is the true value of a parameter" means that "the observation x has distribution P_θ"). Hence, the following definition.

A statistic $T : (X, \mathcal{A}) \to (Y, \mathcal{B})$ is said to be **sufficient** if, for a given value of the statistic T, the distribution of observations x does not depend on θ, i.e., more precisely, for any $A \in \mathcal{A}$,

$$P_\theta \{x \in A \mid T(x)\}$$

does not depend on θ (and, therefore, does not contain any information about θ). The following theorem supplies a convenient method to find a sufficient statistic.

Factorization theorem. Let L be a likelihood function. A statistic $T : (X, \mathcal{A}) \to (Y, \mathcal{B})$ is sufficient if and only if there exists an $\mathcal{A}$-measurable, positive semi-definite function h on X, and $\mathcal{B}$-measurable, positive semi-definite function g_θ on Y, so that

$$L(x \mid \theta) = g_\theta(T(x)) \cdot h(x) \text{ for all } \theta \in \Theta,\ x \in X.$$

Remark 1. The following anticipated statement (obtained from the factorization theorem) can be taken as another definition of a sufficient statistic. A statistic T is sufficient if and only if, for any *a priori* distribution of the parameter θ, the *a posteriori* distribution depends on the observation x through $T(x)$.

Remark 2. The following (equivalent) definition of a sufficient statistic is also interesting. A statistic T is said to be sufficient if there exists a function Q: $\mathcal{X} \to [0, 1]$ such that

$$P_\theta(x \in A) = P_\theta\{T(x) \in T(A)\} \cdot Q(A) \text{ for all } A \in \mathcal{A}.$$

We now prove the factorization theorem only for the case where the observation ξ assumes a finite or countable number of values (X being finite or countable), and $L(x \mid \theta) = P_\theta(\xi = x)$.

If T is a sufficient statistic, $x \in X$, and $T(x) = t$, then

$$\begin{aligned} L(x \mid \theta) &= P_\theta(\xi = x) = P_\theta(\xi = x,\ T(\xi) = t) \\ &= P_\theta\{T(\xi) = t\} \cdot P_\theta\{\xi = x \mid T(\xi) = t\} \\ &= g_\theta(t) \cdot h(x) = g_\theta(T(x)) \cdot h(x), \end{aligned}$$

$g_\theta(t) = P_\theta\{T(\xi) = t\}$, and $h(x) = P_\theta(\xi = x \mid T(\xi) = t)$ not depending on θ, since T is a sufficient statistic.

If, however, the likelihood function L is of the form

$$L(x \mid \theta) = g_\theta(T(x)) \cdot h(x),$$

then, if $T(x) = t$, and $P_\theta\{T(\xi) = t\} > 0$, we obtain

$$\begin{aligned} P_\theta(\xi = x \mid T(\xi) = t) &= \frac{P_\theta(\xi = x,\ T(\xi) = t)}{P_\theta(T(\xi) = t)} \\ &= \frac{P_\theta(\xi = x)}{P_\theta(T(\xi) = t)} = \frac{P_\theta(\xi = x)}{\sum\limits_{y:T(y)=t} P_\theta(\xi = y)} \\ &= \frac{g_\theta(t) \cdot h(x)}{\sum\limits_{y:T(y)=t} g_\theta(t) \cdot h(y)} = \frac{h(x)}{\sum\limits_{y \in T^{-1}(t)} h(y)}, \end{aligned}$$

and the latter expression does not obviously depend on θ.

§ 4. Complete Statistic

A statistic T is said to be (boundedly) **complete** if, for any (bounded) numerical statistic $f(T)$,

$$E_\theta f(T(x)) = 0 \ \forall\, \theta \Rightarrow f(T(x)) = 0 \ \mathcal{P}\text{—a.e.}$$

Consequently,

$$P_\theta\{f(T(x)) = 0\} = 1 \ \forall\, \theta. \tag{1}$$

Theorem. Let a statistical structure $(X, \mathcal{A}, \mathcal{P} = \{P_\theta,\ \theta \in \Theta\})$ admit a likelihood function L of the form

$$\frac{dP_\theta}{d\mu}(x) = L(x \mid \theta) = c(\theta) \cdot h(x) \cdot \exp\left\{\sum_{k=1}^{m} \gamma_k(\theta)\, T_k(x)\right\}$$

(such a family of distributions P_θ is said to be exponential), and there exists a subset $\Theta_0 \subset \Theta$ such that the image of the mapping

$$\Theta_0 \ni \theta \to \gamma(\theta) = \{\gamma_1(\theta), \ldots, \gamma_m(\theta)\} \in R^m$$

contains at least one point along with a certain neighbourhood. Then the statistic $T(x) = \{T_1(x), \ldots, T_m(x)\}$ is complete (and sufficient).

Proof. We may assume that $h(x) \equiv 1$, and $\gamma(\Theta_0)$ contains the rectangle

$$R = \{(\gamma_1, \ldots, \gamma_m) : -c < \gamma_k < c,\ k = 1, \ldots, m\}, \quad c > 0.$$

Let

$$E_\theta f(T) = 0 \ \forall\, \theta \in \Theta_0 \tag{2}$$

for a certain statistic $f(T)$. Put $f(t) = f^+(t) - f^-(t)$, $f^\pm(t) \geqslant 0$, and denote by λ the image of measure μ under the mapping $T : (X, \mathcal{A}) \to (R^m, \mathbf{B}^m)$, i.e., $\lambda = \mu \circ T^{-1}$, or $\lambda(B) = \mu(T^{-1}(B))$ for all $B \in \mathbf{B}^\pm$. Then, for all $\gamma \in R$,

$$\int e^{(\gamma,\, t)} f^+(t)\, \lambda(dt) = \int e^{(\gamma,\, t)} f^-(t)\, \lambda(dt),$$

where $(\gamma, t) = \sum\limits_1^m \gamma_k t_k$. In particular,

$$\int f^+(t)\, \lambda(dt) = \int f^-(t)\, \lambda(dt);$$

meanwhile, without loss of generality, the latter integrals can be assumed to be equal to unity. Putting

$$P^\pm(B) = \int_B f^\pm(t)\, \lambda(dt), \quad B \in \mathbf{B}^m,$$

we obtain that P^+ and P^- are probability measures on $(R^m, \mathbf{B}^m)$, with

$$\int e^{(\gamma, t)} P^+(dt) = \int e^{(\gamma, t)} P^-(dt) \quad \forall\, \gamma \in R^m,$$

from which it follows that these integrals are defined in the strip

$$\{\gamma = u + iv:\ u \in R^m,\ v \in R^m\}$$

of the m-dimensional complex plane; moreover, the above integrals are analytic functions of γ, and, by the analytic extension theorem, coincide in the indicated strip. In particular,

$$\int_{R^n} e^{i(v, t)} P^+(dt) = \int_{R^m} e^{i(v, t)} P^-(dt), \quad v \in R^m.$$

These integrals are the CF's of the distributions P^+ and P^-, respectively, and it follows from the coincidence of these CF's that $P^+ = P^-$; therefore, $f^+(t) = f^-(t)$ a.e. with respect to measure λ, i.e., $f(t) = 0$ λ-a.e. Thus, equality (2) implies (1).

§ 5. Free Statistic

A set $A \in \mathcal{A}$ is said to be **free** (with respect to a family $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ of probability measures on $(X, \mathcal{A})$ if $P_\theta(A)$ does not depend on $\theta \in \Theta$. A statistic T: $(X, \mathcal{A}) \to (Y, \mathcal{B})$ is said to be **free** if the distribution of this statistic does not depend on $\theta \in \Theta$, i.e., $\{x : T(x) \in B\}$ is a free set for all $B \in \mathcal{B}$.

Remark. It is intuitively clear that a free statistic does not contain any information about the true value of parameter θ. On the contrary, a sufficient statistic contains all the information (as much information as in the observation itself) about θ. A sufficient (and boundedly complete) statistic T and a free statistic U often complete each other in the sense that the mapping $x \to \{T(x),\ U(x)\}$ is bijective (T and U being independent as RV's).

Basu theorem. Let $T(x)$ be a sufficient, boundedly complete statistic, and $U(x)$ a free statistic. Then the statistics $T(x)$ and $U(x)$ are independent.

Proof. Since the statistic $T : (X, \mathcal{A}) \to (Y, \mathcal{B})$ is sufficient, and $U : (X, \mathcal{A}) \to (Z, \mathbf{C})$ free, then, for any $C \in \mathbf{C}$ (see also §§ 21, 22, Ch. 2),

$$P_\theta\{U(x) \in C \mid T(x)\} - P_\theta\{U(x) \in C\} = g(T(x))$$

is a free bounded statistic, while

$$E_\theta g(T(x)) = 0 \ \forall\, \theta.$$

Therefore, due to the bounded completeness of the statistic $T(x)$, we have

$$P_\theta\{U(x) \in C \mid T(x)\} = P_\theta\{U(x) \in C\}\ P_\theta\text{-a.e.}\ \forall\, \theta,$$

which implies the independence of the RV's $U(x)$ and $T(x)$ in the probability space $(X, \mathcal{A}, P_\theta)$ for all θ, because, when $B \in \mathcal{B}$, we get

$$\begin{aligned} P_\theta\{U(x) \in C,\ T(x) \in B\} &= E_\theta 1_{\{T(x) \in B\}} \cdot P_\theta\{U(x) \in C \mid T(x)\} \\ &= E_\theta 1_{\{T(x) \in B\}} \cdot P_\theta\{U(x) \in C\} \\ &= P_\theta\{U(x) \in C\} \cdot E_\theta 1_{\{T(x) \in B\}} \\ &= P_\theta\{U(x) \in C\} \cdot P_\theta\{T(x) \in B\}. \end{aligned}$$

Example. Let $x = (x_1, \ldots, x_n)$, the RV's $x_1, \ldots, x_n$ be independent, and each have distribution $N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$. Put

$$T(x) = (\bar{x}, S^2),\ \bar{x} = \frac{1}{n}\sum_1^n x_k, \quad S^2 = \sum_1^n (x_k - \bar{x})^2, \quad S > 0 \text{ a.c.},$$

$$U(x) = \left\{\frac{x_1 - \bar{x}}{S}, \ldots, \frac{\bar{x}_n - x}{S}\right\}.$$

Then the statistic $T(x)$ is complete and sufficient (see the theorem in § 4), whereas $U(x)$ is free (its distribution does not depend on a shift μ and scale σ, i.e., on θ); therefore, by the Basu theorem, the statistics $T(x)$ and $U(x)$ are independent. Note, in addition, that the statistic $U(x)$ assumes values on the intersection of the unit sphere $u_1^2 + \ldots + u_n^2 = 1$ with the plane $u_1 + \ldots + u_n = 0$, while the mapping $x \to \{T(x),\ U(x)\}$ is bijective a.e.

§ 6. Sufficient Condition for Existence of Decision Rule with Uniformly Minimum Risk

1°. Let $(X, \mathcal{A}, \mathcal{P} = \{P_\theta,\ \theta \in \Theta\})$ be a statistical structure, $T(x)$ a sufficient statistic, D a convex set (of possible decisions) in a vector space, and $l(d \mid \theta)$ a loss function convex downward with respect to $d \in D$ for each $\theta \in \Theta$.

We associate each DR δ: $X \to D$ with a DR δ^*: $X \to D$ by putting

(1) $\quad \delta^*(x) = g(T(x)),\ g(T) = E_\theta(\delta(x) \mid T(x) = T) = E_{\theta T}\delta(x).$

It is clear that the latter expression does not depend on θ, since $T(x)$ is a sufficient statistic. It is assumed that δ is integrable. Then

(2) $\quad E_\theta\delta^*(x) = E_\theta\delta(x)$, and $R(\delta^* \mid \theta) \leqslant R(\delta \mid \theta)\ \forall\, \theta.$

The latter statement follows from the Jensen inequality. Note that δ^* depends on x through $T(x)$. Thus, the operation of averaging a DR with respect to a sufficient statistic forms a DR in using which the risk may only decrease.

2°. A DR $\delta : X \to D$ is called an **unbiased estimator** (UE) of a function $\tau : \Theta \to D$ if $E_\theta \delta(x) = (\tau(\theta)\ \forall\ \theta$. It is thereby assumed that δ is an integrable statistic. We denote by $\Delta = \Delta_\tau$ the class of UE's of the function τ. Moreover, we assume that the sufficient statistic $T(x)$ is complete, and $\Delta \neq \varnothing$. Then there exists a DR $\delta^* \in \Delta$ such that

(a) $R(\delta^* \mid \theta) \leqslant R(\delta \mid \theta)\ \forall\ \theta \in \Theta,\ \delta \in \Delta$,

(b) δ^* depends on x through $T(x)$, and

(c) δ^* does not depend on a concrete form of the loss function.

In fact, we select $\delta \in \Delta$, and construct δ^* according to (1). Due to (2), it suffices to show that δ^* does not depend $\mathcal{P}$-a.c. on $\delta \in \Delta$. The latter statement follows from $E_\theta [g_1(T(x)) - g_2(T(x))] = 0\ \forall\ \theta$ implying $g_1(T(x)) = g_2(T(x))$ $\mathcal{P}$-a.c.

3°. As a corollary, we obtain that any (square-integrable) numerical function of a complete sufficient statistic is a ($\mathcal{P}$-a.c. unique) UE (for its ME) with uniformly minimum quadratic risk (i.e., minimum variance).

4°. **Definition.** A family $\mathcal{P}^* = \{P_x^*,\ x \in X\}$ of probability measures on $(X, \mathcal{A})$ is said to be **conjugate** of a family $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ if, for any $A \in \mathcal{A}$,

$$\delta^*(x) = P_x^*(A)$$

is a **minimum variance unbiased estimator** (MVUE) of the function $\tau(\theta) = P_\theta(A)$.

Theorem 1. Let $(X, \mathcal{A})$ be a Borel space (e.g., separable, metric; see § 22, Ch. 2), and there exist a complete sufficient statistic. Then there exists a family $\mathcal{P}^*$ conjugate of $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$.

Proof. Let $T(x)$ be a complete sufficient statistic. By Theorem 1, § 22, Ch. 2, for each $\theta \in \Theta$, there exists regular conditional probability relative to the RV $T(x)$. In our case, $(\Omega, S, P) = (X, \mathcal{A}, P_\theta)$ is a probability space, $\eta(x) = x$ an RV with values in the Borel space, and $\mathcal{F}$ the σ-algebra generated by the RV $T(x)$. It remains to put

$$\delta^*(x) = P_x^*(A) = P_\theta(A \mid T)(x) \text{ for } A \in \mathcal{A}. \tag{3}$$

Note that the latter expression does not depend on θ, since $T(x)$ is a sufficient statistic. It now remains to see that

$$E_\theta\, \delta^*(x) = P_\theta(A) = \tau(\theta),$$

and that δ^* depends on X through $T(x)$, where $T(x)$ is a complete statistic (see 3°).

Theorem 2. Let the conditions of Theorem 1 be fulfilled, the family $\mathcal{P}^*$ conjugate of $\mathcal{P}$, and $g(x)$ a $\mathcal{P}$-(square-)integrable statistic. Put

$$\tau(\theta) = E_\theta g(x) = (P_\theta,\ g)\ \forall\ \theta.$$

Then

$$\delta^*(x) = \int_X g(u)\, P_x^*(du) = (P_x^*,\ g)$$

is an MVUE of the function $\tau(\theta)$.

Proof. According to the definition of P_x^* (see (3)), $\delta^*(x)$ can be written in the form

$$\delta^*(x) = \int_X g(u)\, P_\theta(du \mid T)(x),$$

which, by Theorem 2, § 22, Ch. 2, equals

$$\delta^*(x) = E_\theta(g \mid T)(x),$$

whence $E_\theta \delta^*(x) = E_\theta g(x) = \tau(\theta)$. It remains to notice that $\delta^*(x)$ depends on x through $T(x)$, where T is a complete statistic.

Remark 1. Under the conditions of Theorem 1, the estimator $\delta^*(x)$ can be conveniently found by formula (1), where $\delta(x) = 1_A(x)$ is a UE $\tau_\theta = P_\theta(A)$.

Remark 2. To make applications more convenient, Theorems 1 and 2 may be represented (strengthened) as follows.

Theorem 1′. Let $(X, \mathcal{A}, \mathcal{P} = \{P_\theta,\ \theta \in \Theta\})$ be a statistical structure, $T(x)$ a sufficient statistic, and $\alpha(x)$ a certain statistic with values in a Borel space $(Y, \mathcal{B})$. Then there exists such a family $\{Q_x:\ x \in X\}$ of probability measures on $(Y, \mathcal{B})$ that

(1) for any $B \in \mathcal{B}$, the estimator $\delta^*(x) = Q_x(B)$ is a UE of the function $\tau_\theta = P_\theta\{\alpha(x) \in B\}$,

(2) $Q_x(B) = P_\theta(\alpha \in B \mid T)$ $\mathcal{P}$-a.c.,

and

(3) $Q_x(B)$ depends on x through $T(x)$.

It is clear that if T is a complete statistic, then $\delta^*(x)$ is the MVUE for $\tau(\theta)$. Theorem 2′ is formulated similarly.

Example 1. Let $x = (x_1, \ldots, x_n)$, where the RRV's $x_1, \ldots, x_n$ are i.i.d. according to the DF F_θ. Assume that there exists a complete, sufficient statistic $T(x)$, and find the MVUE of the function $\tau(\theta) = F_\theta(t)$, where t is a certain number. The estimator

$$\delta(x) = \begin{cases} 1, & x_1 < t \\ 0, & x_1 \geqslant t \end{cases}$$

is unbiased; therefore, the required estimator is $\delta^*(x) = g(T(x))$, where $g(T) = P_\theta\{x_1 < t \mid T(x) = T\}$.

Putting

$$\delta^*(x) = F_x^*(t),$$

we obtain that the function $t \to F_x^*(t)$ is a DF, and, for any number t, the estimator $F_x^*(t)$ is the MVUE for $F_\theta(t)$. If g is an F_θ-integrable

function for all θ, then

$$\delta^*(x) = \int_{-\infty}^{\infty} g(t)\, dF_x^*(t)$$

is the MVUE of the function

$$\tau(\theta) = \int_{-\infty}^{\infty} g(t)\, dF_\theta(t).$$

Example 2. Let $x = (x_1, \ldots, x_m)$, $y = (y_1, \ldots, y_n)$, the RV's $x_1, \ldots, x_m, y_1, \ldots, y_n$ independent, F_a the DF of the RV's $x_1, \ldots, x_m$, G_b the DF of the RV's $y_1, \ldots, y_n$, the parameters a and b unknown, and $\theta = (a, b)$. Assume that $T_1 = T_1(x_1, \ldots, x_m)$, $T_2 = T_2(y_1, \ldots, y_n)$, and $T = (T_1, T_2)$ are complete, sufficient statistics with respect to the parameters a, b and $\theta = (a, b)$, respectively. For any number t, we denote, similarly to the previous example, the MVUE for $F_a(t)$ by $F_x^*(t)$, and that for $G_b(t)$ by $G_y^*(t)$. Let P_θ and $P_{x,y}^*$ be two probability measures (on the plane) generated by the DF's $F_a(t_1) \cdot G_b(t_2)$ and $F_x^*(t_1) \cdot G_y^*(t_2)$, respectively. Then, for any measurable set A on the plane, $\delta^*(x, y) = P_{x,y}^*(A)$ is the MVUE for $\tau(\theta) = P_\theta(A)$. In particular, if two independent RV's ξ and η have distributions F_a and G_b, while independent ξ^* and η^* have distributions F_x^* and G_x^*, respectively, then

$$\delta^*(x, y) = P_{x,y}^*(\xi^* < \eta^*)$$

is the MVUE for

$$\tau(\theta) = P_\theta(\xi < \eta).$$

Remark. The previous example is an illustration of the following general statement. Let a statistical structure $(X, \mathcal{A}, \mathcal{P} = \{P_\theta, \theta \in \Theta\})$ be the product of s statistical structures

$$(X_k, \mathcal{A}_k, \mathcal{P}_k = \{P_{\theta_k}, \theta_k \in \Theta_k\}) \quad k = 1, \ldots, s,$$

where $\mathcal{P}$ is the family of probability measures $P_\theta = P_{\theta_1} \times \ldots \times P_{\theta_s}$, $\theta = (\theta_1, \ldots, \theta_s) \in \Theta = \Theta_1 \times \ldots \times \Theta_s$ on the measurable spaces

$$(X, \mathcal{A}) = (X_1, \mathcal{A}_1) \times \ldots \times (X_s, \mathcal{A}_s).$$

Assume that each of the structures possesses a complete, sufficient statistic. If $\mathcal{P}_k^* = \{P_{x_k}^*, x_k \in X_k\}$, $k = 1, \ldots, s$, is a family of probability measures on $(X_k, \mathcal{A}_k)$, conjugate of $\mathcal{P}_k$, then the family $\mathcal{P}^*$ of probability measures $P_x^* = P_{x_1}^* \times \ldots \times P_{x_s}^*$, $x = (x_1, \ldots, x_s) \in X = X_1 \times \ldots \times X_s$, on the measurable space $(X, \mathcal{A})$ is conjugate of the family $\mathcal{P}$. In particular, if an RV ξ has a distribution $P_\theta \in \mathcal{P}$, and the RV ξ^* $P_x \in \mathcal{P}^*$, then

$$\delta(x) = E_x^* g(\xi^*)$$

is the MVUE of the function

$$\tau(\theta) = E_\theta g(\xi),$$

where E_x^* is averaging with respect to measure P_x^*, and g a $\mathcal{P}$-(square-) integrable function.

Chapter 10

Point Estimation: Cramér-Rao Theory

§ 1. MVUE

Let $(X, \mathcal{A}, \mathcal{P} = \{P_\theta, \theta \in \Theta\})$ be a statistical structure, and τ a certain mapping of Θ into R^m. A statistic t with values in R^m is called an **estimator** of the function τ.

An estimator t of a function τ is said to be **unbiased** if

$$E_\theta t(x) = \tau(\theta) \; \forall \theta,$$

where E_θ is the ME operator, provided the RV x has a distribution P_θ.

As a rule, the requirement of unbiasedness alone does not distinguish an estimator t uniquely. Another requirement for the estimators is that of variance minimum. Let Δ be a class of UE's. An estimator $t \in \Delta$ is called a **minimum variance unbiased estimator** if

$$\operatorname{var}_\theta t \leqslant \operatorname{var}_\theta t' \; \forall \theta \in \Theta, \quad t' \in \Delta.$$

Generally speaking, these requirements already distinguish an estimator t uniquely if it exists at all, which is not always valid, since the variance must be minimum uniformly with respect to θ. This circumstance is, perhaps, the most serious argument against such strong requirements. Note that if $m > 1$, and $t = (t_1, \ldots, t_m)'$, $\tau = (\tau_1, \ldots, \tau_m)'$, then

$$\operatorname{var}_\theta t = E_\theta (t - \tau)(t - \tau)' = \{\operatorname{cov}(t_i, t_j)\},$$

and the inequality $A \leqslant B$ for two matrices A and B of the same order means that the matrix $B - A$ is positive semi-definite ($B - A \geqslant 0$).

In this chapter, we shall study the MVUE. We shall show that such an estimator is unique, indicate a method of finding it, and prove the positive definite property for estimators of this kind (their dependence on an observation x through $T(x)$, where T is a sufficient statistic).

Example 1 (*where a* UE *exists*). Let $x = (x_1, \ldots, x_n)$, where the RV's $x_1, \ldots, x_n$ are i.i.d. according to $N(\mu, \sigma^2)$. Put $\theta = (\mu, \sigma^2)$, $\Theta = R^1 \times R^1_+$. Then

$$t_1(x) = \bar{x} = \frac{1}{n}\sum_1^n x_k \text{ is the UE of } \tau_1(\theta) = \mu,$$

$$t_2(x) = \frac{1}{n-1}\sum_1^n (x_k - \bar{x})^2 \text{ is the UE of } \tau_2(\theta) = \sigma^2.$$

Example 2 (*where a* UE *does not exist*). Let x have the Poisson distribution P_θ, i.e.,

$$P_\theta(x = k) = \frac{\theta^k}{k!} e^{-\theta}, k = 0, 1, 2, \ldots,$$

and $\Theta = (0, \infty)$. Then there exists no UE of the function $\tau(\theta) = 1/\theta$. In fact, if t is a UE of τ, then

$$E_\theta t(x) = \sum_{k \geqslant 0} t(k) \frac{\theta^k}{k!} e^{-\theta} = 1/\theta \quad \forall\, \theta > 0,$$

or

$$\sum_{k \geqslant 0} \frac{t(k)}{k!} \theta^k = e^\theta/\theta \quad \forall\, \theta > 0,$$

which is impossible.

Example 3 (*where a* UE *is absurd*). Let x have the "truncated" Poisson distribution with parameter θ, i.e.,

$$P_\theta(x = k) = \frac{\theta^k}{k!} e^{-\theta}[1 - e^{-\theta}]^{-1}, \quad k = 1, 2, \ldots,$$

$\Theta = (0, \infty)$, and $\tau(\theta) = 1 - e^{-\theta}$. If t is a UE of τ, then

$$E_\theta t(x) = \sum_{k \geqslant 1} t(k) \frac{\theta^k}{k!} e^{-\theta}[1 - e^{-\theta}]^{-1} = 1 - e^{-\theta} \quad \forall\, \theta > 0,$$

or

$$\sum_{k \geqslant 1} \frac{t(k)}{k!} \theta^k = e^\theta[1 - e^{-\theta}]^2 = e^\theta + e^{-\theta} - 2 = \sum_{m \geqslant 1} \frac{2}{(2m)!} \theta^{2m}$$

for all $\theta > 0$, from which it follows that

$$t(x) = \begin{cases} 0 \text{ if } x \text{ is even} \\ 2 \text{ if } x \text{ is odd} \end{cases}$$

It is the unique UE of the function τ. Note that $0 < \tau(\theta) < 1$.

§ 2. MVUE Uniqueness

Theorem. Let t_1 and t_2 be two UE's of a function τ with minimum variance. Then $t_1(x) = t_2(x)$ $\mathcal{P}$-a.c., i.e.,

$$P_\theta\{t_1(x) = t_2(x)\} = 1 \quad \forall\, \theta$$

(in other words, the statistics t_1 and t_2 are equivalent).

Proof. For $\theta \in \Theta$, we put

$$t_3 = \frac{1}{2}(t_1 + t_2), \quad v = \operatorname{var}_\theta t_1 = \operatorname{var}_\theta t_2.$$

Then t_3 is a UE of the function τ, and $v \leqslant \operatorname{var}_\theta t_3$. On the other hand,

$$\operatorname{var}_\theta t_3 = \frac{1}{4}[\operatorname{var}_\theta t_1 + \operatorname{var}_\theta t_2 + 2\operatorname{cov}_\theta(t_1, t_2)] \leqslant v,$$

since, according to the Cauchy-Schwartz inequality,

$$\operatorname{cov}_\theta(t_1, t_2) \leqslant [\operatorname{var}_\theta t_1 \cdot \operatorname{var}_\theta t_2]^{1/2} = v.$$

Therefore, $\operatorname{var}_\theta t_3 = v$, and $\operatorname{cov}_\theta(t_1, t_2) = v$, whence

$$\operatorname{var}_\theta(t_1 - t_2) = \operatorname{var}_\theta t_1 + \operatorname{var}_\theta t_2 - 2\operatorname{cov}_\theta(t_1, t_2) = 0,$$

which is equivalent to the equality

$$P_\theta(t_1 = t_2) = 1.$$

Remark. If t_1 and t_2 are two vector-valued MVUE's of a function τ, then, for any vector $c \in R^m$, the estimators (c, t_1) and (c, t_2) are the MVUE's of the function (c, τ); therefore, as it follows from what has already been proved, $(c, t_1) = (c, t_2)$ $\mathcal{P}$-a.c., thus implying that $t_1 = t_2$ $\mathcal{P}$-a.c.

§ 3. Cramér-Rao Inequality

Assume that a statistical structure $(X, \mathcal{A}, \mathcal{P} = \{P_\theta, \theta \in \Theta\})$ admits a likelihood function

$$L(x|\theta) = \frac{dP_\theta}{d\mu}(x)$$

with respect to a certain measure μ on $(X, \mathcal{A})$. Consider also a UE t of a function τ.

In this section, we assume that Θ is an open set on the straight line R^1, and t, τ two numerical functions. Besides, we assume that the functions L, t and τ satisfy the **regularity condition,** viz.,

(1) the set $\{x : L(x \mid \theta) > 0\}$ does not depend on θ, in which case we may assume without loss of generality that $L(x \mid \theta) > 0$ $\forall\, x$ and θ,

(2) the function $L(x \mid \theta)$ is differentiable with respect to θ, while

$$\frac{d}{d\theta}\int_X L(x|\theta)\,\mu(dx) = \int_X \frac{\partial}{\partial\theta} L(x|\theta)\,\mu(dx),$$

$$\frac{d}{d\theta}\int_X t(x)\,L(x|\theta)\,\mu(dx) = \int_X t(x)\,\frac{\partial}{\partial\theta} L(x|\theta)\,\mu(dx), \quad \text{and}$$

(3) the function τ is differentiable.

Theorem. If t is a UE of a function τ with finite second moment, then

$$\operatorname{var}_\theta t \geqslant \frac{[\tau'(\theta)]^2}{E_\theta\left(\frac{\partial}{\partial\theta}\ln L\right)^2} \quad \forall\,\theta.$$

This inequality turns into an equality if and only if

$$\frac{\partial}{\partial\theta}\ln L = A(\theta)\,[t(x)-\tau(\theta)] \quad (\mu\text{-a.c.})$$

for a certain numerical function $A(\theta)$.

Proof. We have

(1) $\int L(x \mid \theta)\,\mu(dx) = 1\ \forall\,\theta$, and

(2) $\int t(x)\,L(x \mid \theta)\,\mu(dx) = \tau(\theta)\ \forall\,\theta$.

Differentiate (1) and (2) with respect to θ. We get

$$0 = \int \frac{\partial L}{\partial\theta}\,d\mu = \int\left(\frac{1}{L}\,\frac{\partial L}{\partial\theta}\right)L\,d\mu = \int \frac{\partial \ln L}{\partial\theta}\,L\,d\mu,$$

$$\tau'(\theta) = \int t\,\frac{\partial L}{\partial\theta}\,d\mu = \int t\,\frac{\partial \ln L}{\partial\theta}\,L\,d\mu,$$

whence

(3) $$\tau'(\theta) = \int [t(x)-\tau(\theta)]\,\frac{\partial \ln L}{\partial\theta}\,L\,d\mu.$$

Apply the Cauchy-Schwartz inequality to the latter integral, representing $L = \sqrt{L}\cdot\sqrt{L}$. We obtain

$$[\tau'(\theta)]^2 \leqslant \int [t(x)-\tau(\theta)]^2\,L\,d\mu \cdot \int\left(\frac{\partial \ln L}{\partial\theta}\right)^2 L\,d\mu$$

$$= \operatorname{var}_\theta t \cdot E_\theta\left(\frac{\partial \ln L}{\partial\theta}\right)^2,$$

the equality occurring if and only if the factors (as functions of x) $\frac{\partial \ln L}{\partial\theta}$ and $t-\tau(\theta)$ are equal (μ-a.c.) up to constant multiple (which may depend on θ, and which we denote by $A(\theta)$).

Remark 1. If an LF L is twice differentiable with respect to θ, and

$$\frac{d^2}{d\theta^2}\int L(x|\theta)\,\mu(dx) = \int \frac{\partial^2}{\partial\theta^2}\,L(x|\theta)\,\mu(dx),$$

then it follows from (1) that

$$0 = \int \frac{\partial^2 L}{\partial\theta^2}\,d\mu = \int \frac{\partial}{\partial\theta}\left(\frac{\partial \ln L}{\partial\theta}L\right)d\mu$$

$$= \int \frac{\partial^2 \ln L}{\partial\theta^2}\,L\,d\mu + \int\left(\frac{\partial \ln L}{\partial\theta}\right)^2 L\,d\mu,$$

i.e.,

$$E_\theta\left(\frac{\partial \ln L}{\partial\theta}\right)^2 = -E_\theta\,\frac{\partial^2 \ln L}{\partial\theta^2};$$

therefore,

$$\operatorname{var}_\theta t \geqslant -\frac{[\tau'(\theta)]^2}{E_\theta\,\frac{\partial^2 \ln L}{\partial\theta^2}} \quad \forall\,\theta.$$

Remark 2. Let us specify the conditions under which

(4) $$\frac{d}{d\theta}\int_X f(x,\theta)\,\mu(dx) = \int_X \frac{\partial}{\partial\theta}\,f(x,\theta)\,\mu(dx).$$

Let f be a numerical function defined on $X\times\Theta$, where Θ is an open set on the straight line R^1, and μ a measure on a measurable space $(X, \mathcal{A})$. Assume that

(1) for any $\theta\in\Theta$, the function $x\to f(x,\theta)$ is integrable with respect to measure μ,

(2) for any $x\in X$, the function $\theta\to f(x,\theta)$ is differentiable, and

(3) the derivative $\frac{d}{d\theta}f$ satisfies the inequality

$$\left|\frac{\partial}{\partial\theta}\,f(x,\theta)\right| \leqslant g(x),$$

where g is a function integrable with respect to measure μ. Then the function

$$\int_X f(x,\theta)\,\mu(dx)$$

of θ is differentiable, and (4) fulfilled. The proof is the same as for Theorem 1, § 7, Ch. 6.

§ 4. Efficient Estimators

The right-hand side of the Cramér-Rao inequality is (for each θ) a lower bound of a UE variance. UE's for which this bound is attained are said to be **efficient.** It is clear that an efficient estimator (EE) is an MVUE; therefore, it is unique. The following theorem indicates a method of obtaining EE's (and, hence, MVUE's).

Theorem. For a UE t of a function τ to be efficient, it is necessary and sufficient that

(1) $$\frac{\partial}{\partial\theta}\ln L = A(\theta)\,[t(x)-\tau(\theta)] \quad (\mu\text{-a.c.}),$$

where A is a certain function only of θ, and

(2) $$\operatorname{var}_\theta t(x) = \frac{\tau'(\theta)}{A(\theta)}.$$

This statement immediately follows from the theorem in § 3, and equality (2) is obtained if we substitute the expression from (1) for $\frac{\partial}{\partial\theta}\ln L$ in formula (3). We certainly assume that the conditions of the theorem in § 3 are fulfilled.

Example 1. Let $x = (x_1, \ldots, x_n)$, the observations $x_1, \ldots, x_n$ be independent, and each have the normal distribution $N(\theta, \sigma^2)$ with known variance σ^2. The density of the distribution of each observation has the form

$$f(y|\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y-\theta)^2}{2\sigma^2}\right\}.$$

In our case,

$$L(x|\theta) = \prod_1^n f(x_k|\theta) = \sigma^{-n}(2\pi)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_1^n (x_k-\theta)^2\right\},$$

$$\frac{\partial}{\partial\theta}\ln L = \frac{n}{\sigma^2}(\bar{x}-\theta), \text{ where } \bar{x} = \frac{1}{n}\sum_1^n x_k.$$

This equality coincides with (1) if

$$A(\theta) = \frac{n}{\sigma^2}, \quad t(x) = \bar{x}, \quad \tau(\theta) = \theta.$$

We obtain from (2) that $\mathrm{var}_\theta t(x) = \sigma^2/n$. Thus, $\bar{x}$ is the MVUE of the parameter θ with variance σ^2/n.

Example 2. Now, let $x = (x_1, \ldots, x_n)$, the RV's $x_1, \ldots, x_n$ be independent, and each have the distribution $N(\mu, \theta^2)$ with known shift μ, in which case

$$L(x|\theta) = \prod_1^n \frac{1}{\theta\sqrt{2\pi}} \exp\left\{-\frac{(x_k-\mu)^2}{2\theta^2}\right\}$$

$$= \theta^{-n}(2\pi)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\theta^2}\sum_1^n (x_k-\mu)^2\right\},$$

$$\frac{\partial}{\partial\theta}\ln L = \frac{n}{\theta^3}\left[\frac{1}{n}\sum_1^n (x_k-\mu)^2 - \theta^2\right]. \tag{3}$$

The latter expression coincides with (1) if we put

$$A(\theta) = \frac{n}{\theta^3}, \quad t(x) = \frac{1}{n}\sum_1^n (x_k-\mu)^2, \quad \tau(\theta) = \theta^2.$$

Then, according to (2),

$$\mathrm{var}_\theta\, t(x) = \frac{2\theta^4}{n}.$$

Remark. It is clear that if t is an EE (of a function τ), then $at + b$ is also an EE (of the function $a\tau + b$), where a and b are two numbers. There are no other EE's. In particular, we can derive from formula (3) that, with the data of Example 2, there exists an EE of the square of the parameter θ, but there exists no EE of the parameter θ itself.

Example 3. If $x = (x_1, \ldots, x_n)$, the RV's $x_1, \ldots, x_n$ are independent, and each has the Poisson distribution, then, putting

$$f(y|\theta) = \frac{\theta^y}{y!} e^{-\theta}, \quad y = 0, 1, 2, \ldots,$$

we obtain

$$L(x|\theta) = \prod_1^n f(x_k|\theta) = \frac{\theta^{x_1+\ldots+x_n}}{x_1!\ldots x_n!} e^{-n\theta},$$

$$\frac{\partial}{\partial\theta}\ln L = \frac{n}{\theta}(\bar{x}-\theta), \quad \bar{x} = \frac{1}{n}(x_1+\ldots+x_n);$$

therefore, $\bar{x}$ is the EE of the parameter θ with variance θ/n.

Example 4. If RV's $x_1, \ldots, x_n$ are independent, and each has the Cauchy distribution with density

$$f(y|\theta) = \frac{1}{\pi}\frac{1}{1+(y-\theta)^2}, \quad y, \theta \in R^1,$$

then

$$\frac{\partial}{\partial\theta}\ln L = 2\sum_{i=1}^n \frac{x_i-\theta}{1+(x_i-\theta)^2}.$$

This equality may have form (1); therefore, there exists no EE (no function of the parameter θ).

Example 5. Let us estimate θ in the binomial distribution $\mathrm{Bi}(\theta, n)$ for which

$$L_x(x|\theta) = {}^nC_x\theta^x(1-\theta)^{n-x}, \quad x = 0, 1, \ldots, n.$$

We find

$$\frac{\partial}{\partial\theta}\ln L = \frac{n}{\theta(1-\theta)}\left(\frac{x}{n}-\theta\right).$$

Therefore, x/n is the MVUE of the parameter θ with variance $\frac{\theta(1-\theta)}{n}$.

Example 6. Let $x = (x_1, \ldots, x_n)$, the RV's $x_1, \ldots, x_n$ be i.i.d. with density of the form

$$f(y \mid \theta) = \exp\{A(\theta)\cdot B(y) + C(\theta) + D(y)\}$$

(such a family of distributions is said to be exponential). If the regularity condition is fulfilled, then

$$L(x|\theta)=\prod_1^n f(x_k|\theta)=\exp\{A(\theta)\cdot\sum_1^n B(x_k)+nC(\theta)+\sum_1^n D(x_k)\},$$

$$\frac{\partial}{\partial\theta}\ln L=nA'(\theta)\left[\frac{1}{n}\sum_1^n B(x_k)+\frac{C'(\theta)}{A'(\theta)}\right].$$

Putting

$$t(x)=\frac{1}{n}\sum_1^n B(x_k),\quad \tau(\theta)=-\frac{C'(\theta)}{A'(\theta)},$$

we obtain that t is the EE of the function τ.

§ 5. Cramér-Rao Inequality Improvement. Bhattacharya Inequality

We can find a better (i.e., greater) lower bound for a UE variance than the right-hand side of the Cramér-Rao inequality in the cases where the latter is not attained, i.e., where there exists no EE. The basic condition for the variance lower bound in the Cramér-Rao inequality to be attained consists in the existence of an estimator t, for which $t-\tau(\theta)$ is a linear function (for each θ) of

$$\frac{\partial}{\partial\theta}\ln L=\frac{1}{L}\cdot\frac{\partial L}{\partial\theta}.$$

However, even if such an estimator does not exist, there may exist an estimator for which $t-\tau(\theta)$ is (for each θ) a linear combination of functions

$$\frac{1}{L}\cdot\frac{\partial L}{\partial\theta},\quad \frac{1}{L}\cdot\frac{\partial^2 L}{\partial\theta^2},\quad\ldots,\quad \frac{1}{L}\cdot\frac{\partial^s L}{\partial\theta^s},$$

which leads to the following result of Bhattacharya. Put

$$L^{(k)}=\frac{\partial^k L}{\partial\theta^k},\quad \tau^{(k)}=\frac{d^k}{d\theta^k}\tau(\theta),$$

and assume that the regularity condition similar to that introduced in § 3 is fulfilled, but the differentiation operation with respect to θ is now used s times.

Theorem. Let t be a UE of τ. Then

$$\operatorname{var}_\theta t\geqslant\sum_{i,j=1}^s a_{ij}c_ic_j \tag{1}$$

where

$$a_{ij}=E_\theta\left(\frac{L^{(i)}}{L}\cdot\frac{L^{(j)}}{L}\right),$$

and the coefficients $c_i=c_i(\theta)$ are determined by the system

$$\sum_{j=1}^s a_{ij}c_j=\tau^{(i)},\quad i=1,\ldots,s. \tag{2}$$

If the matrix $A=\{a_{ij}\}$ is invertible, and $A^{-1}=\{a^{ij}\}$, then

$$\operatorname{var}_\theta t\geqslant\sum_{i,j=1}^s a^{ij}\tau^{(i)}\tau^{(j)}. \tag{3}$$

The latter expression is called the **Bhattacharya inequality.** Inequalities (1) and (3) turn into equalities if and only if

$$t-\tau(\theta)=\sum_{i=1}^s c_i\frac{L^{(i)}}{L} \tag{4}$$

for certain $c_i=c_i(\theta)$.

Proof. A. Lemma. Let $\alpha,\alpha_1,\ldots,\alpha_s$ be elements of a certain linear space with scalar product and norm $|\alpha|=(\alpha,\alpha)^{1/2}$. Then

$$|\alpha|^2\geqslant\sum_{i,j=1}^s(\alpha_i,\alpha_j)c_ic_j, \tag{5}$$

where the numbers $c_1,\ldots,c_s$ satisfy the system of equations

$$\sum_{j=1}^s(\alpha_i,\alpha_j)c_j=(\alpha,\alpha_i)\ \forall\, i, \tag{6}$$

with equality in (5) occurring if and only if

$$\alpha=\sum_1^s c_i\alpha_i \tag{7}$$

for certain numbers $c_1,\ldots,c_s$.

In fact, let L be the space of linear combinations of the elements $\alpha_1,\ldots,\alpha_s$, and

$$\beta=\operatorname{proj}_L\alpha=\sum_1^s c_i\alpha_i$$

the projection of α onto L. Then

$$|\alpha|^2\geqslant|\beta|^2=(\sum_1^s c_i\alpha_i,\ \sum_1^s c_j\alpha_j)=\sum(\alpha_i,\alpha_j)c_ic_j$$

which gives (5). Meanwhile, $\alpha-\beta\perp\alpha_i\ \forall\, i$, thus yielding (6). Besides, $|\alpha|=|\beta|$ if and only if $\alpha\in L$, which gives (7).

B. For a fixed value of the parameter θ, we consider the linear space of RRV's with scalar product $(\xi,\eta)=E_\theta(\xi\cdot\eta)$. Put

$$\alpha=t-\tau(\theta),\quad \alpha_i=L^{(i)}/L,\quad i=1,\ldots,s.$$

Then $|\alpha|^2=\operatorname{var}_\theta t$, $(\alpha_i,\alpha_j)=a_{ij}$, $(\alpha,\alpha_i)=\tau^{(i)}$, in which case (1), (2) and (4) coincide with (5)-(7) of the lemma. We now verify

that $(\alpha, \alpha_i) = \tau^{(i)}$. We obtain from the equalities

$$\int_X L(x|\theta)\,\mu(dx) = 1, \quad \int_X t(x)\,L(x|\theta)\,\mu(dx) = \tau(\theta)$$

that

$$\int L^{(i)}\,d\mu = \int \frac{L^{(i)}}{L}\cdot L\,d\mu = 0, \quad \int t\cdot\frac{L^{(i)}}{L}\cdot L\,d\mu = \tau^{(i)},$$

respectively, or $E_\theta\alpha_i = 0$, $E_\theta t\cdot\alpha_i = \tau^{(i)}$, which yields

$$(\alpha, \alpha_i) = E_\theta t\,\alpha_i - \tau(\theta)\cdot E_\theta\alpha_i = \tau^{(i)}.$$

C. It remains to prove that the right-hand sides of (1) and (3) coincide (when the matrix A is invertible). Put

$$c = (c_1, \ldots, c_s)', \quad \tau = (\tau^{(1)}, \ldots, \tau^{(s)})'.$$

Then (2) is written in the form $Ac = \tau$ and inequality (1) as $\mathrm{var}_\theta\, t \geqslant (Ac, c)$. It remains to notice that

$$(Ac, c) = (A^{-1}\tau, \tau) = \sum_{i,j=1}^{s} a^{ij}\tau^{(i)}\tau^{(j)}$$

§ 6. Multidimensional Analogue of Cramér-Rao Inequality

Earlier, we have considered the case where θ is a numerical parameter. We will now assume that θ is a vector of dimension r with real components, i.e., $\theta = (\theta_1, \ldots, \theta_r) \in \Theta$, and Θ is an open set in R^r. Consider an estimator $t(x)$ of a function $\tau(\Theta)$, and assume that the regularity condition is fulfilled (see § 3, where differentiation means partial differentiation, and also Remark 2, § 3).

Theorem. Let $t(x)$ be a UE of a numerical function $\tau(\theta) = \tau(\theta_1, \ldots, \theta_r)$. Then

$$\mathrm{var}_\theta\, t \geqslant \sum_{i,j=1}^{r} a_{ij}c_ic_j, \tag{1}$$

where

$$a_{ij} = a_{ij}(\theta) = E_\theta\left(\frac{\partial \ln L}{\partial\theta_i}\cdot\frac{\partial \ln L}{\partial\theta_j}\right),$$

and the coefficients $c_i = c_i(\theta)$ satisfy the system of equations

$$\sum_{j=1}^{r} a_{ij}c_j = \frac{\partial\tau}{\partial\theta_i}, \quad i = 1, \ldots, r.$$

If, however, the matrix $A = \{a_{ij}\}$ is invertible, and $A = \{a^{ij}\}$, then

$$\mathrm{var}_\alpha\, t \geqslant \sum_{i,j=1}^{r} a^{ij}\frac{\partial\tau}{\partial\theta_i}\cdot\frac{\partial\tau}{\partial\theta_j}, \tag{2}$$

with the equality in (1) and (2) occurring if and only if

$$t - \tau(\theta) = \sum_{i=1}^{r} c_i(\theta)\,\frac{\partial \ln L}{\partial\theta_i}, \tag{3}$$

where c_i may depend only on θ.

Proof follows from the lemma of the previous section if we introduce scalar product $(\xi, \eta) \in E_\theta(\xi\cdot\eta)$ in the linear space of RRV's with finite second moments, and put

$$\alpha = t - \tau(\theta), \quad \alpha_i = \frac{\partial \ln L}{\partial\theta_i}.$$

Then

$$|\alpha|^2 = \mathrm{var}_\theta\, t, \quad (\alpha_i\alpha_j) = a_{ij},$$

$$(\alpha, \alpha_i) = E_\theta(t-\tau)\frac{\partial \ln L}{\partial\theta_i} = E_\theta t\,\frac{\partial \ln L}{\partial\theta_i} = \int_X t\,\frac{\partial L}{\partial\theta_i}\,d\mu = \frac{\partial\tau}{\partial\theta_i}.$$

Remark. If the LF $L(x \mid \theta)$ is twice differentiable with respect to θ, then

$$a_{ij} = E_\theta\left(\frac{\partial \ln L}{\partial\theta_i}\cdot\frac{\partial \ln L}{\partial\theta_j}\right) = -E_\theta\frac{\partial^2 \ln L}{\partial\theta_i\,\partial\theta_j}.$$

It is sometimes easier to employ the latter expression in computations (see Remark 1, § 3).

Corollary (Information Inequality). Consider the case where an estimator t and the function τ (of the parameter) are vector functions. Put

$$t(x) = (t_1(x), \ldots, t_m(x))', \quad \tau(\theta) = (\tau_1(\theta), \ldots, \tau_m(\theta))',$$

$$\mathrm{var}_\theta\, t = \{\mathrm{cov}_\theta(t_i, t_j)\}, \quad \frac{\partial\tau}{\partial\theta} = \left\{\frac{\partial\tau_i}{\partial\theta_j}\right\}.$$

We show that if t is a UE of the function τ, then

$$\mathrm{var}_\theta\, t \geqslant \left(\frac{\partial\tau}{\partial\theta}\right)' A^{-1}\left(\frac{\partial\tau}{\partial\theta}\right), \tag{4}$$

where we still have

$$A = \{a_{ij}\}, \quad a_{ij} = E_\theta\left(\frac{\partial \ln L}{\partial\theta_i}\cdot\frac{\partial \ln L}{\partial\theta_j}\right),$$

and the inequality $C \geqslant D$ for matrices (of the same order) means that the matrix $C - D$ is positive semi-definite.

In fact, for $c = (c_1, \ldots, c_m)' \in R^m$, we obtain that $\hat{t}(x) = (c, t(x))$ is a UE of the function $\hat{\tau}(\theta) = (c, \tau(\theta))$; therefore, equality (2) written in the form

$$\mathrm{var}_\theta\,\hat{t} \geqslant \left(A^{-1}\frac{\partial\hat{\tau}}{\partial\theta}, \frac{\partial\hat{\tau}}{\partial\theta}\right), \quad \text{where} \quad \frac{\partial\hat{\tau}}{\partial\theta} = \left(\frac{\partial\hat{\tau}}{\partial\theta_1}, \ldots, \frac{\partial\hat{\tau}}{\partial\theta_r}\right)'$$

holds. However,

$$\operatorname{var}_\theta \hat{t} = \sum_{i,j=1}^{m} c_i c_j \operatorname{cov}_\theta (t_i, t_j) = (Vc, c), \text{ where } V = \operatorname{var}_\theta t,$$

$$\frac{\partial \hat{\tau}}{\partial \theta} = \sum_{1}^{m} c_i \frac{\partial \tau_i}{\partial \theta} = \left(\frac{\partial \tau}{\partial \theta}, c\right);$$

hence,

$$(Vc, c) \geqslant \left(A^{-1} \frac{\partial \tau}{\partial \theta} c, \frac{\partial \tau}{\partial \theta} c\right) = (Bc, c),$$

where we have put

$$B = \left(\frac{\partial \tau}{\partial \theta}\right)' A^{-1} \frac{\partial \tau}{\partial \theta}.$$

Thus,

$$(Vc, c) \geqslant (Bc, c), \text{ or } ((V - B) c, c) \geqslant 0 \quad \forall\, c \in R^m,$$

which just means that the matrix $V - B$ is positive semi-definite.

Note that, in the case $\tau(\theta) = \theta \quad \forall \theta \in \Theta$,

$$\operatorname{var}_\theta t \geqslant A^{-1}. \tag{5}$$

This is called the **information inequality**, and the matrix A the **Fisher information matrix.**

Example. Let $x = (x_1, \ldots, x_n)$, the RV's $x_1, \ldots, x_n$ be independent, and each have distribution density of the form

$$f(y \mid \theta) = \exp \{A(\theta) \cdot B(y) + C(\theta) + D(y)\}, \quad \theta = (\theta_1, \ldots, \theta_r).$$

Then the estimator

$$t(x) = \frac{r}{n} \sum_{1}^{n} B(x_k)$$

is the MVUE of the function $\tau(\theta) = -\sum_{1}^{r} \frac{\partial C}{\partial \theta_i} \left(\frac{\partial A}{\partial \theta_i}\right)^{-1}$. In fact, (3) is valid in this case if we put

$$c_i(\theta) = \left(n \frac{\partial A}{\partial \theta_i}\right)^{-1}.$$

§ 7. Multidimensional Analogue of Bhattacharya Inequality

Note one of useful applications of the Cramér-Rao, Bhattacharya, and multidimensional analogue of the Cramér-Rao inequalities, viz. that they yield a sufficient condition for an estimate $t(x)$ to be an MVUE (it consists in the necessity of turning into an equality at least one of the inequalities for all θ). It consecutively follows from these

inequalities that, for an estimator $t(x)$ to be an MVUE, it is sufficient that the estimator $t(x)$ of a function $\tau(\theta)$ should be of the form

$$\frac{\partial}{\partial \theta} \ln L = A(\theta)\,[t - \tau(\theta)] \tag{1}$$

for the case where θ is a numerical parameter, or, more generally,

$$t - \tau(\theta) = \sum_{1}^{s} c_i(\theta) \frac{L^{(i)}}{L}, \quad L^{(i)} = \frac{\partial^i L}{\partial \theta^i}, \text{ and}$$

$$t - \tau(\theta) = \sum_{i=1} c_i(\theta) \frac{\partial \ln L}{\partial \theta_i} \tag{2}$$

for the case where θ is a vector parameter of dimension r, or, more generally (see the muldtidimensional analogue of the Bhattacharya inequality below),

$$t - \tau(\theta) = \frac{1}{L} \left\{ \sum \frac{\partial L}{\partial \theta_i} \cdot c_i(\theta) + \sum \frac{\partial^2 L}{\partial \theta_i\, \partial \theta_j} \cdot c_{ij}(\theta) + \ldots + \sum \frac{\partial^s L}{\partial \theta_{i_1} \ldots \partial \theta_{i_s}} c_{i_1 \ldots i_s}(\theta) \right\}.$$

Before stating the Bhattacharya inequality for the case of a vector parameter $\theta = (\theta_1, \ldots, \theta_r) \in \Theta$, where Θ is an open set in R^r, we introduce the following simple notation. We define partial differentiation operators by putting

$$D_i = \frac{\partial}{\partial \theta_i}, \quad D^\alpha = D_1^{\alpha_1} \ldots D_r^{\alpha_r}, \quad |\alpha| = \alpha_1 + \ldots + \alpha_r$$

for any vector $\alpha = (\alpha_1, \ldots, \alpha_r)$ whose components are non-negative integers.

The regularity condition is assumed to be valid, it being similar to that employed in § 3, with the only difference that the (partial) differentiation operation is applied s times.

Theorem. Let $t(x)$ be a UE of a function $\tau(\theta) = \tau(\theta_1, \ldots, \theta_r)$. Then

$$\operatorname{var}_\theta t \geqslant \sum_{\substack{1 \leqslant |\alpha| \leqslant s \\ 1 \leqslant |\beta| \leqslant s}} a_{\alpha\beta} c_\alpha c_\beta,$$

where

$$a_{\alpha\beta} = E_\theta \left(\frac{L^{(\alpha)}}{L} \cdot \frac{L^{(\beta)}}{L}\right), \quad L^{(\alpha)} = D^\alpha L,$$

and the coefficients $c_\alpha = c_\alpha(\theta)$ satisfy the system of equations

$$\sum_{1 \leqslant |\beta| \leqslant s} a_{\alpha\beta} c_\beta = \tau^{(\alpha)}, \quad 1 \leqslant |\alpha| \leqslant s, \quad \tau^{(\alpha)} = D^\alpha \tau.$$

Meanwhile, the inequality turns into an equality if and only if

$$t-\tau(\theta)=\sum_{1\leqslant|\alpha|\leqslant s} c_\alpha(\theta)\frac{L^{(\alpha)}}{L}$$

for certain coefficients c_α which may depend only on θ.

Proof is similar to that performed for the case of a vector parameter θ (see § 6).

Example 1. Let $x=(x_1, \ldots, x_n)$, the RV's $x_1, \ldots, x_n$ be independent, and each have distribution $N(\theta_1, \theta_2^2)$, with $\theta=(\theta_1, \theta_2)$, $-\infty<\theta_1<\infty$, $\theta_2>0$,

$$L(x|\theta)=\theta_2^{-n}(2\pi)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\theta_2^2}\sum_{k=1}^{n}(x_k-\theta_1)^2\right\}.$$

Put

$$t(x)=\frac{1}{n-1}\sum_1^n(x_k-\bar{x})^2,\quad \bar{x}=\frac{1}{n}\sum_1^n x_k,\quad \tau(\theta)=\theta_2^2.$$

Then

$$t-\tau(\theta)=\frac{1}{L}\left\{\frac{\theta_2^3}{n-1}\cdot\frac{\partial L}{\partial\theta_2}-\frac{\theta_2^4}{n(n-1)}\cdot\frac{\partial^2 L}{\partial\theta_1^2}\right\}.$$

It follows from the theorem and this representation that $t(x)$ is the MVUE of the function $\tau(\theta)$. By the way, the theorems of the previous sections, where the Cramér-Rao and Bhattacharya inequalities were investigated, do not permit us to make such a conclusion. To derive this simple corollary, a stronger statement is required. It is perhaps sometimes a serious argument to assert that efficient estimation theory is inefficient.

Example 2 (*continued*). If we do not restrict ourselves to the class of UE's, then we can find a "better" estimator of the variance θ_2^2 than that offered in the previous example. Consider the quadratic risk of an estimator $\delta(x)$ of the function $\tau(\theta)=\theta_2^2$, viz.,

$$R(\delta\mid\theta)=E_\theta[\delta(x)-\tau(\theta)]^2.$$

Then

$$R(t^\star\mid\theta)<R(t\mid\theta)=\operatorname{var}_\theta t\ \forall\,\theta$$

is valid for the estimator

$$t^\star(x)=\frac{1}{n+1}\sum_1^n(x_k-\bar{x})^2=\frac{1}{n+1}S.$$

Note that the offered estimator $t^\star$ has minimum quadratic risk in the class of estimators of the form $c\cdot S$, $c\in R^1$.

§ 8. Expressing MVUE in Terms of Sufficient Statistic

Let T be a sufficient statistic, and t an MVUE. It is natural to expect that $t(x)$ depends on x through $T(x)$, i.e., $t(x)=h(T(x))$ for a certain function h. We prove that it is really so, for which we shall make use of the results of §§ 21, 22, Ch. 2, regarding the conditional expectation and conditional probability.

Since $E(\eta\mid\xi)=g(\xi)$ for a certain (measurable) function g, we obtain

$$E_\theta(t(x)\mid T(x))=h(T(x))$$

for a certain (measurable) function h. Since $T(x)$ is a sufficient statistic, $t_1(x)=h(T(x))$ does not depend on θ, and, because $E\{E(\eta\mid\xi)\}=E\eta$, we get

$$E_\theta t_1(x)=E_\theta\{E_\theta(t(x)\mid T(x))\}=E_\theta t(x)=\tau(\theta),$$

i.e., both t_1 and t are UE's of the function $\tau(\theta)$. Due to the uniqueness of an MVUE, it suffices to show that $\operatorname{var}_\theta t\geqslant\operatorname{var}_\theta t_1$. Putting

$$E_\theta(t(x)\mid T(x))=E_{\theta,T(x)}\,t(x),$$

and making use of the Jensen inequality, we have

$$\begin{aligned}\operatorname{var}_\theta t&=E_\theta[t(x)-\tau(\theta)]^2=E_\theta\{E_{\theta,T(x)}[t(x)-\tau(\theta)]^2\}\\&\geqslant E_\theta[E_{\theta,T(x)}t(x)-\tau(\theta)]^2\\&=E_\theta[t_1(x)-\tau(\theta)]^2=\operatorname{var}_\theta t_1.\end{aligned}$$

Chapter 11

Point Estimation: Maximum Likelihood Method

§ 1. Maximum Likelihood Estimator

A useful method for deriving reasonable estimators is the maximum likelihood method, according to which an estimator $\hat{\theta}=\hat{\theta}(x)$ is obtained from the condition

$$L(x\mid\hat{\theta})\geqslant L(x\mid\theta)\ \forall\,\theta,$$

i.e.,

$$L(x\mid\hat{\theta})=\max_\theta L(x\mid\theta).$$

Such an estimator is called a **maximum likelihood estimator (MLE).**

If $\Theta \subset R^s$ and, for any $x \in X$, the LF $L(x \mid \theta)$ is differentiable with respect to θ, while attaining its maximum at an interior point (which belongs to Θ along with a certain neighbourhood), then the MLE $\hat{\theta}$ satisfies the condition

$$\frac{\partial}{\partial \theta} L = 0, \quad \text{or} \quad \frac{\partial}{\partial \theta} \ln L = 0 \text{ when } \theta = \hat{\theta},$$

where $\frac{\partial}{\partial \theta} L$ is the vector

$$\frac{\partial}{\partial \theta} L = \left\{ \frac{\partial L}{\partial \theta_1}, \ \ldots, \ \frac{\partial L}{\partial \theta_s} \right\} \text{ for } \theta = (\theta_1, \ \ldots, \ \theta_s) \in \Theta.$$

Note, in addition, that if the EE of the parameter θ exists, then it can be obtained by the maximum likelihood method, since (when $s = 1$),

$$\frac{\partial \ln L}{\partial \theta} = A(\theta)\,[t(x) - \theta].$$

In the general case (where a statistical structure admits an LF), the MLE $\hat{\theta}(x)$ depends on x through $T(x)$, where T is a sufficient statistic; this follows from the factorization

$$L(x \mid \theta) = g_\theta(T(x))\, h(x)$$

and definition of an MLE.

Example 1. Assume that $x = (x_1, \ldots, x_n)$, and the RV's $x_1, \ldots, x_n$ are i.i.d. according to $N(\mu, \sigma^2)$, in which case

$$L(x \mid \theta) = \prod_{k=1}^{n} \frac{1}{\sigma \sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2} (x_k - \mu)^2 \right\}$$

$$= \sigma^{-n} (2\pi)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_1^n (x_k - \mu)^2 \right\}.$$

Let the parameters μ and $\sigma > 0$ be unknown. Put $\theta = (\theta_1, \theta_2) = (\mu, \sigma) \in \Theta = R^1 \times R^1_+$ and find the MLE $\hat{\theta} = (\hat{\mu}, \hat{\sigma})$. Since $L = 0$ holds on the boundary of Θ, maximum is attained at an interior point, and

$$\frac{\partial}{\partial \mu} \ln L = \frac{1}{\sigma^2} \sum_1^n (x_k - \mu) = 0,$$

$$\frac{\partial}{\partial \sigma} \ln L = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_1^n (x_k - \mu)^2 = 0$$

for $\theta = \hat{\theta}$, whence

$$\hat{\mu} = \bar{x} = \frac{1}{n}(x_1 + \ldots + x_n), \quad \hat{\sigma}^2 = \frac{1}{n} \sum_1^n (x_k - \bar{x})^2.$$

Example 2. Assume now that $x = (x_1, \ldots, x_n)$, and the random vectors $x_1, \ldots, x_n$ are i.i.d. according to $N_r(\mu, A)$, in which case $\theta = (\mu, A)$, and

$$L(x \mid \theta) = \sum_{k=1}^{n} \frac{1}{(2\pi)^{r/2} |A|^{1/2}} \exp\left\{ -\frac{1}{2} (A^{-1}(x_k - \mu), \ x_k - \mu) \right\}$$

$$= (2\pi)^{-\frac{nr}{2}} |A|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \sum_1^n (A^{-1}(x_k - \mu), \ x_k - \mu) \right\}.$$

Since

$$L(x|\theta) = (2\pi)^{-\frac{nr}{2}} |A|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \sum (A^{-1}(x_k - \bar{x}), \ x_k - \bar{x}) \right.$$

$$\left. -\frac{n}{2} (A^{-1}(\bar{x} - \mu), \ \bar{x} - \mu) \right\},$$

the determination of the maximum of the LF $L(x \mid \theta)$ from the vector $\mu \in R^r$ is equivalent to that of the expression

$$(A^{-1}(\bar{x} - \mu), \ \bar{x} - \mu)$$

from μ. Since A (and, therefore, A^{-1}) is a positive definite matrix, the minimum of the latter expression is attained at the point

$$\mu = \hat{\mu} = \bar{x} = \frac{1}{n}(x_1 + \ldots + x_n).$$

Put

$$\hat{A} = \frac{1}{n} \sum_1^n (x_k - \bar{x})(x_k - \bar{x})',$$

then (taking into account that $(a, b) = \operatorname{tr}(ab')$ for column vectors a and b)

$$\sum_1^n (A^{-1}(x_k - \bar{x}), \ x_k - \bar{x}) = \sum_1^n \operatorname{tr}(A^{-1} \cdot (x_k - \bar{x}) \cdot (x_k - \bar{x})')$$

$$= \operatorname{tr}\left(A^{-1} \cdot \sum_1^n (x_k - \bar{x})(x_k - \bar{x})'\right) = n \operatorname{tr}(A^{-1}\hat{A}),$$

and

$$L(x|\hat{\mu}, A) = (2\pi)^{-\frac{nr}{2}} |A|^{-\frac{n}{2}} \exp\left\{ -\frac{n}{2} \operatorname{tr}(A^{-1}\hat{A}) \right\}.$$

We now show that the estimator $\hat{\theta} = (\hat{\mu}, \hat{A})$ is an MLE. We have seen that, for any positive definite matrix A,

$$L(x \mid \mu, A) \leqslant L(x \mid \hat{\mu}, A)$$

is valid for all $\mu \in R^r$. It now suffices to see that

$$L(x \mid \hat{\mu}, A) \leqslant L(x \mid \hat{\mu}, \hat{A})$$

for any positive definite matrix A. The latter inequality is equivalent to the following (we have to take the logarithm of each side), viz.,

$$-\frac{n}{2} \ln |A| - \frac{n}{2} \operatorname{tr}(A^{-1}\hat{A}) \leqslant -\frac{n}{2} \ln |\hat{A}| - \frac{n}{2} \operatorname{tr} I_r,$$

where I_r is the unit matrix of order $r \times r$, or

$$\ln |A^{-1}\hat{A}| - \operatorname{tr}(A^{-1}\hat{A}) + r \leqslant 0.$$

Since A and $\hat{A}$ are positive definite matrices (a.e. if $n > r$), they can be represented in the form

$$A = B^2, \quad \hat{A} = \hat{B}^2,$$

where B and $\hat{B}$ are symmetric matrices. Put $C = (B^{-1}\hat{B}) \times (B^{-1}\hat{B})'$. Then the matrix C is positive definite. Besides, if f is a numerical function of a matrix of the form $f(A) = |A|$ or $f(A) = \operatorname{tr} A$, then $f(AB) = f(BA)$; therefore,

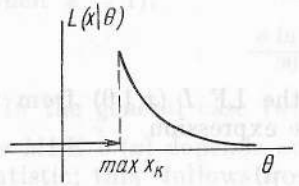

Fig. 14

$$f(A^{-1}\hat{A}) = f(B^{-1}B^{-1}\hat{B}\hat{B}) = f(B^{-1}\hat{B}\hat{B}B^{-1}) = f(B^{-1}\hat{B}\,(B^{-1}\hat{B})') = f(C).$$

Hence, the above inequality is equivalent to

$$\ln |C| - \operatorname{tr} C + r \leqslant 0$$

or

$$\sum_{i=1}^{r} (\ln \lambda_i - \lambda_i + 1) \leqslant 0,$$

where $\lambda_1, \ldots, \lambda_r$ are the eigenvalues of the matrix C, since

$$|C| = \lambda_1 \ldots \lambda_r, \quad \operatorname{tr} C = \lambda_1 + \ldots + \lambda_r.$$

It remains to notice that each addend in the above sum is non-positive, because $\lambda_i > 0$, and $\ln x \leqslant x - 1$ for all $x > 0$.

Note that

$$L(x|\hat{\mu}, \hat{A}) = (2\pi e)^{-\frac{nr}{2}} |\hat{A}|^{-\frac{n}{2}}.$$

Example 3. If $x = (x_1, \ldots, x_n)$, the RV's $x_1, \ldots, x_n$ are i.i.d. on $[0, \theta]$, $\theta > 0$, then

$$L(x|\theta) = \prod_{k=1}^{n} \frac{1}{\theta} 1_{[0,\,\theta]} x_k = \theta^{-n} 1_{[0,\,\theta]}(\max x_k) \cdot 1_{[0,\,\infty)}(\min x_k),$$

where, as usual, 1_A is the indicator of the set A (see Fig. 14). It is clear that the MLE is $\hat{\theta}(x) = \max(x_1, \ldots, x_n)$.

§ 2. Consistent Estimators

Let, as before, an unknown parameter θ which we cannot observe be seeked. Suppose, however, that RV's with distribution depending on θ are observed. Let $x_1, x_2, \ldots$ be the sequence of the results of observations. Denote by $t_n = t_n(x_1, \ldots, x_n)$ an estimator of the parameter θ, constructed from the first n observations $x_1, \ldots, x_n$. There has been no necessity before to employ the subscript n in the notation of an estimator t_n, since the number n of observations was fixed. If, however, the quantity of observations increases, then the estimators t_n, in case they are reasonable, should tend to the parameter θ under estimation. These arguments lead to the following definition.

An estimator t_n depending on the number of observations n is said to be consistent (CE) of a parameter θ if the sequence $\{t_n\}$ converges to θ. As a rule, convergence means convergence in probability in this case, i.e.,

$$P_\theta \{|t_n - \theta| \geqslant \varepsilon\} \to 0 \text{ as } n \to \infty$$

for any $\varepsilon > 0$ (and any θ!).

In the general case, if not the parameter itself is estimated, but a certain function of this parameter, say, $\tau(\theta)$, then the estimator t_n is called a CE of the function $\tau(\theta)$ if, for any θ, the sequence $t_n \to \tau(\theta)$ in probability (i.e., P_θ).

Remark. Though the requirement that an estimator should be consistent is desirable for any estimation rule, this requirement is not at all related to any property if the sample size is fixed. Consistency is a property of a sequence of estimators (for increasing sample size). In this sense, we had better say that the sequence $\{t_n\}$ of estimators, and not an estimator t_n, is consistent. Thus, if t_n is a CE, then

$$t'_n = \begin{cases} \theta_0 & \text{when } n \leqslant 2^{64} \\ t_n & \text{when } n > 2^{64} \end{cases}$$

is also a CE. Certainly, if a certain method of obtaining estimators leads to inconsistent estimators, then the consistency of this method is doubtful.

§ 3. MLE Consistency

Assume that, for each integer $n \geqslant 1$, a LF $L(x \mid \theta) = L_n(x \mid \theta)$ has the form

$$L_n(x \mid \theta) = f(x_1 \mid \theta) \ldots f(x_n \mid \theta), \quad x = (x_1, x_2, \ldots),$$

which means that the RV's $x_1, x_2, \ldots$, are i.i.d. with density $f(y \mid \theta)$. The family of distributions F_θ of the RV x_1 for $\theta \in \Theta$ is dominated by a certain measure μ, and $f(y \mid \theta) = (dF_\theta/d\mu)(y)$.

Case 1: the set Θ is finite. We introduce the following assumptions.

A1. The set $Y = \{y : f(y \mid \theta) > 0\}$ does not depend on $\theta \in \Theta$. We can thereby assume that the density $f(y \mid \theta)$ is defined only on the rectangle $Y \times \Theta$, and takes positive values.

A2. $\theta_1 = \theta_2 \Leftrightarrow f(y \mid \theta_1) = f(y \mid \theta_2)$ a.e. (with respect to measure μ).

A3. There exists

$$E_{\theta_0} \ln \frac{f(x_1 \mid \theta_0)}{f(x_1 \mid \theta)} = \varphi(\theta \mid \theta_0) = \varphi(\theta),$$

where θ_0 is a fixed element from Θ.

A4. For each integer $n \geqslant 1$, the MLE $\hat{\theta}_n$ is unique a.c.

Theorem 1.

$$P_\theta \{\hat{\theta}_n = \theta_0\} \to 1 \quad \text{as} \quad n \to \infty.$$

Corollary 1. If A3 is valid for all $\theta_0 \in \Theta$, then

$$P_\theta \{\hat{\theta}_n = \theta\} \to 1 \quad \text{as} \quad n \to \infty \ \forall\, \theta \in \Theta,$$

i.e., the MLE $\hat{\theta}_n$ is consistent.

Proof. 1°. Let us verify that $\varphi(\theta) \geqslant 0 \ \forall\, \theta \in \Theta$, and $\varphi(\theta) = 0 \Leftrightarrow \theta = \theta_0$, which follows from the Jensen inequality applied to the strictly convex function $g(x) = -\ln x$ on $(0, \infty)$, taking into account A2, and

$$E\xi = 1, \text{ where } E = E_{\theta_0}, \text{ and } \xi = f(x_1 \mid \theta)/f(x_1 \mid \theta_0),$$

since

$$E\xi = \int_Y \frac{f(y \mid \theta)}{f(y \mid \theta_0)} \cdot f(y \mid \theta_0)\, \mu(dy) = 1.$$

2°. Put

$$\xi_n(\theta) = \frac{1}{n} \sum_1^n \ln \frac{f(x_k \mid \theta_0)}{f(x_k \mid \theta)}.$$

Since $E\xi_n(\theta) = \varphi(\theta)$, according to the Khinchine law of large numbers, $\xi_n(\theta) \to \varphi(\theta)$ in probability. If, however, $\theta \neq \theta_0$, and, therefore, $\varphi(\theta) > 0$, then $\forall \alpha > 0\ \exists\, N$ such that $P\{\xi_n(\theta) > 0\} \geqslant 1 - \alpha$ for $n \geqslant N$, where $P = P_{\theta_0}$. Since Θ consists of a finite number of elements, we can assume that it is valid for all $\theta \neq \theta_0$.

3°. Put $A_{n\theta} = \{\xi_n(\theta) > 0\}$. Since $\xi_n(\theta) > 0 \ \forall\, \theta \neq \theta_0 \Leftrightarrow L_n(x \mid \theta_0) > L_n(x \mid \theta) \ \forall\, \theta \neq \theta_0 \Leftrightarrow \theta \neq \theta_0$, $P(\hat{\theta}_n = \theta_0) = P\left(\bigcup_{\theta \neq \theta_0} A_{n\theta}\right)$, whence, putting $|\Theta| = m$, and taking into account (2°), we obtain

$$P(\hat{\theta}_n \neq \theta_0) = P\left(\bigcup_{\theta \neq \theta_0} A^c_{n\theta}\right) \leqslant \sum_{\theta \neq \theta_0} P(A^c_{n\theta}) < (m-1)\,\alpha$$

for $n \geqslant N$, which gives $P(\hat{\theta}_n \neq \theta_0) \to 0$ due to the arbitrariness of the choice of $\alpha > 0$.

Case 2: the set Θ is infinite. We make the following assumptions.

A1. The density $f(y \mid \theta)$ is defined on the rectangle $Y \times \Theta$, and takes positive values.

A2. Θ is a locally compact, metric space with metric ρ (local compactness means that any point from Θ contains a neighbourhood whose closure is compact).

A3. Y is a compact, metric space.

A4. The function $f(y \mid \theta)$ is continuous on the metric space.

A5. $\theta_1 = \theta_2 \Leftrightarrow f(y \mid \theta_1) = f(y \mid \theta_2)\ \forall\, y$.

A6. There exists

$$E_{\theta_0} \ln \frac{f(x_1 \mid \theta_0)}{f(x_1 \mid \theta)} = \varphi(\theta \mid \theta_0) = \varphi(\theta),$$

and the function $\varphi(\theta)$ is continuous on Θ, where θ_0 is a fixed element from Θ.

A7. The MLE of $\hat{\theta}_n$ is the unique local maximum of the LF $L_n(x \mid \theta)$, which means that if $\varepsilon > 0$, $\theta_0 \in \Theta$, and $L_n(x \mid \theta_0) > L(x \mid \theta)$ for all points, θ of the sphere $\rho(\theta, \theta_0) = \varepsilon$, then $\hat{\theta}_n$ is inside this sphere, i.e., $\rho(\hat{\theta}_n, \theta_0) < \varepsilon$.

Theorem 2. For any $\varepsilon > 0$,

$$P_{\theta_0} \{\rho(\theta_n, \theta_0) < \varepsilon\} \to 1 \text{ as } n \to \infty.$$

Corollary 2. If A6 is valid for all $\theta_0 \in \Theta$, then, for all $\theta \in \Theta$ and $\varepsilon > 0$,

$$P_\theta \{\rho(\hat{\theta}_n, \theta_0) < \varepsilon\} \to 1 \text{ as } n \to \infty$$

holds, i.e., the MLE $\hat{\theta}_n$ is consistent.

Proof. 1°. Just like in the proof of Theorem 1, we verify that $\varphi(\theta) \geqslant 0 \ \forall\, \theta$, and $\varphi(\theta) = 0 \Leftrightarrow \theta = \theta_0$.

2°. Since there exists a neighbourhood of the point $\theta_0 \in \Theta$, whose closure is compact, then there exists $\varepsilon_0 > 0$ such that, for $0 < \varepsilon < \varepsilon_0$, the ball with radius ε and centre at the point θ_0 is contained in this neighbourhood; in particular, the sphere

$$S = \{\theta\colon \rho(\theta, \theta_0) = \varepsilon\}$$

is a compactum. But, according to 1°, the continuous function $\varphi(\theta)$ is positive on this compactum; therefore,

$$\delta = \min_{\theta \in S} \varphi(\theta) > 0.$$

3°. $\forall \delta_1 > 0\ \exists\, \varepsilon_1 > 0$ such that $|\varphi(\theta_1) - \varphi(\theta_2)| < \delta_1$ for $\rho(\theta_1, \theta_2) < \varepsilon_1$, $\theta_1 \in S$, $\theta_2 \in S$.

4°. Put

$$g(y, \theta_1, \theta_2) = \ln \frac{f(y \mid \theta_1)}{f(y \mid \theta_2)}.$$

The numerical function g is continuous on the compact set $Y \times S \times S$; therefore, it is uniformly continuous, and, since $g(y, \theta, \theta) = 0$, we derive, in particular, that $\forall \delta_1 > 0\ \exists\ \varepsilon_1 > 0$ such that

$$| g(y, \theta_1, \theta_2) | < \delta_1$$

for $y \in Y$, $\rho(\theta_1, \theta_2) < \varepsilon_1$, $\theta_1 \in S$, $\theta_2 \in S$. Applied to the function

$$\xi_n(\theta) = \frac{1}{n} \sum_1^n \ln \frac{f(x_k \mid \theta_0)}{f(x_k \mid \theta)},$$

this yields

$$|\xi_n(\theta_2) - \xi_n(\theta_1)| = \left| \frac{1}{n} \sum_1^n g(x_k, \theta_1, \theta_2) \right| < \delta_1$$

under the same conditions on θ_1, θ_2. We can assume that ε_1 is the same in 3° and 4°.

5°. Now, for $\varepsilon_1 > 0$, we put $U_\theta = \{\theta' \in \Theta\colon \rho(\theta', \theta) < \varepsilon_1\}$. The family of open sets U_θ, $\theta \in S$, covers the compact set S; therefore, there exists a finite subfamily of open sets U_θ, $\theta \in S_1 \subset S$, which cover S.

6°. It follows from the law of large numbers in the Khinchine form that $\xi_n(\theta)$ tends to $E_{\theta_0} \xi_n(\theta) = \varphi(\theta)$ in probability as $n \to \infty$. In particular, $\forall\ \delta_1 > 0$, $\alpha > 0$, $\theta_1 \in S\ \exists\ N$ such that $P\{|\xi_n(\theta_1) - \varphi(\theta_1)| > \delta_1\} < \alpha$ when $n > N$, $P = P_{\theta_0}$, from which it follows that $\forall\ \delta_1 > 0$, $\alpha > 0$ and any finite set $S_1 \subset S$, there exists N such that

$$P\{\bigcup_{\theta \in S_1} (|\xi_n(\theta) - \varphi(\theta)| > \delta_1)\} < \alpha \text{ when } n > N,$$

or

$$P\{|\xi_n(\theta) - \varphi(\theta)| \leqslant \delta_1\ \forall\ \theta \in S_1\} \geqslant 1 - \alpha \quad \text{when} \quad n > N.$$

We shall see below that the above inequality is valid if S_1 is replaced by S, and δ_1 by $3\delta_1$.

7°. However, we first have to verify that, for any number $c > 0$, the set $\{|\xi_n(\theta) - \varphi(\theta)| \leqslant c\ \forall\ \theta \in S\}$ is P-measurable, or that

$$\eta_n = \eta_n(x_1, \ldots, x_n) \sup_{\theta \in S} |\xi_n(\theta) - \varphi(\theta)|$$

is an RV, for which we put $\eta_n(\theta) = |\xi_n(\theta) - \varphi(\theta)|$, and notice that, for all $x_1, \ldots, x_n$, the function $\eta_n(\theta)$ is continuous on the sphere S, while the compactum S contains a countable set $S_c \subset S$, everywhere dense in S (since any compactum is a separable space). We obtain

$$\eta_n = \sup_{\theta \in S} \eta_n(\theta) = \sup_{\theta \in S_c} \eta_n(\theta).$$

But the latter quantity is an RV as the supremum of a countable (!) number of RV's.

8°. We first show that $|\xi_n(\theta_1) - \varphi(\theta_1)| \leqslant \delta_1\ \forall\ \theta_1 \in S_1$ implies $|\xi_n(\theta) - \varphi(\theta)| \leqslant 3\delta_1\ \forall\ \theta \in S$. In fact, if $\theta \in S_1$, and $|\xi_n(\theta_1) - \varphi(\theta_1)| \leqslant \delta_1$, then, for $\theta \in U_{\theta_1} \cap S$, we have $\rho(\theta, \theta_1) \leqslant \varepsilon_1$; therefore, according to 4° and 3°,

$$|\xi_n(\theta) - \xi_n(\theta_1)| \leqslant \delta_1, \quad \text{and} \quad |\varphi(\theta) - \varphi(\theta_1)| \leqslant \delta_1,$$

which implies $|\xi_n(\theta) - \varphi(\theta)| \leqslant 3\delta_1$ for $\theta \in U_{\theta_1} \cap S$. It remains to notice that the sets $U_{\theta_1} \cap S$, $\theta_1 \in S_1$ cover (according to 5°) the set S. It follows from what has been proved that

$$P\{\eta_n(\theta) \leqslant 3\delta_1\ \forall\ \theta \in S\} \geqslant P\{\eta_n(\theta) \leqslant \delta_1\ \forall\ \theta \in S_1\},$$

which gives

$$P\{|\xi_n(\theta) - \varphi(\theta)| \leqslant 3\delta_1\ \forall\ \theta \in S\} \geqslant 1 - \alpha \quad \text{when } n > N.$$

9°. In particular, it follows from the implication

$$\eta_n(\theta) \leqslant 3\delta_1 \Rightarrow \xi_n(\theta) \geqslant \varphi(\theta) - 3\delta_1 \geqslant \delta - 3\delta_1 > 0$$

(for $\delta_1 < \delta/3$) that $P\{\xi_n(\theta) > 0\ \forall\ \theta \in S\} \geqslant 1 - \alpha$ when $n > N$.

But

$$\xi_n(\theta) > 0\ \forall\ \theta \in S \Leftrightarrow L_n(x \mid \theta_0) > L_n(x \mid \theta)\ \forall\ \theta \in S \Leftrightarrow \rho(\hat{\theta}_n, \theta_0) < \varepsilon;$$

therefore,

$$P\{\rho(\hat{\theta}_n, \theta_0) < \varepsilon\} \geqslant 1 - \alpha \quad \text{when } n > N.$$

Remark. It can be seen from the proof that, instead of A1 and A4, it is sufficient to assume that the function

$$g(y, \theta_1, \theta_2) = \ln \frac{f(y \mid \theta_1)}{f(y \mid \theta_2)}$$

is defined on $Y \times \Theta \times \Theta$, and continuous on $Y \times \Theta_0 \times \Theta_0$, where Θ_0 is a certain neighbourhood of the point θ_0. In particular, the density $f(y \mid \theta)$ can assume zero value at certain points y, and then the function g must be defined at these points by continuity (if possible).

§ 4. MLE Asymptotic Normality

We now list the conditions under which the MLE $\hat{\theta} = \hat{\theta}_n$ is asymptotically normal, i.e., the distribution of the RV $\sqrt{n}(\hat{\theta}_n - \theta)$ tends to the normal distribution $N(\theta, A_\theta)$.

We again assume that, for any integer $n \geqslant 1$,

$$L(x \mid \theta) = L_n(x \mid \theta) = f(x_1 \mid \theta) \times \ldots \times f(x_n \mid \theta), x = (x_1, x_2, \ldots, x_n).$$

We now make the following assumptions.

A1. The density f is defined on $Y \times \Theta$, and $f(y \mid \theta) > 0$ for all $(y, \theta) \in Y \times \Theta$.

A2. Θ is a convex set in R^s.

A3. The LF attains its global maximum at an interior point (i.e., at a point $\hat{\theta}$ belonging to Θ along with a certain neighbourhood) almost for all x.

A4. The MLE $\hat{\theta} = \hat{\theta}_n$ is consistent.

A5. The function $f(y \mid \theta)$ is twice differentiable a.e. at the point $\theta = \theta_0$ under the integral sign, i.e., there exists

$$\left\{\frac{\partial^2}{\partial\theta_i\,\partial\theta_j} f(y|\theta)\right\}, \quad \theta = (\theta_1, \ldots, \theta_s)',$$

and

$$\frac{\partial}{\partial\theta_i}\int_Y f(y|\theta)\,\mu(dy) = \int_Y \frac{\partial}{\partial\theta_i} f(y|\theta)\,\mu(dy),$$

$$\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\int_Y f(y|\theta)\,\mu(dy) = \int_Y \frac{\partial^2}{\partial\theta_i\,\partial\theta_j} f(y|\theta)\,\mu(dy) \quad \text{for } \theta = \theta_0.$$

A6. If

$$F(x_1|\theta) = \frac{\partial^2}{\partial\theta^2}\ln f(x_1|\theta) = \left\{\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\ln f(x_1|\theta)\right\},$$

then

$$\| F(x_1|\theta) - F(x_1|\theta_0) \| \leqslant c(x_1|\theta_0)\cdot g(\theta|\theta_0),$$

with

$$c(x_1 \mid \theta_0) \geqslant 0, \text{ and } E_{\theta_0} c(x_1 \mid \theta_0) < \infty,$$

whereas the function $g(\theta \mid \theta_0)$ is continuous with respect to θ at the point θ_0, and $g(\theta_0 \mid \theta_0) = 0$.

A7. The matrix $J_0 = J(\theta_0)$ is invertible, where

$$J(\theta) = \left\{E_\theta\left(\frac{\partial \ln f(x_1|\theta)}{\partial\theta_i}\cdot\frac{\partial \ln f(x_1 \mid \theta)}{\partial\theta_j}\right)\right\}.$$

Theorem. The distribution of the RV $\sqrt{n}\,(\hat{\theta}_n - \theta_0)$ tends to the normal distribution $N_s(0, J_0^{-1})$ as $n \to \infty$, i.e.,

$$P_{\theta_0}\{\sqrt{n}\,(\hat{\theta}_n - \theta_0) < u\} \to P(\xi < u) \text{ as } n \to \infty$$

for any $u \in R^s$, where the RV ξ has the distribution $N_s(0, J_0^{-1})$.

Corollary. If A5-A7 are valid for all $\theta_0 \in \Theta$, then

$$P_\theta\{\sqrt{n}\,(\hat{\theta}_n - \theta) < u\} \to P(\xi < u) \text{ as } n \to \infty$$

for all $\theta \in \Theta$ and $u \in R^s$, whereas ξ has the distribution $N(0, J(\theta)^{-1})$.

Proof. 1°. Put

$$g(t) = u(x \mid \theta_t),\ \theta_t = \theta_t(x) = (1-t)\,\theta_0 + t\hat{\theta},\ 0 \leqslant t \leqslant 1,$$

$$u(x|\theta) = \frac{\partial \ln L}{\partial\theta} = \left\{\frac{\partial \ln L}{\partial\theta_1}, \ldots, \frac{\partial \ln L}{\partial\theta_s}\right\}'.$$

Then

$$g(1) = \frac{\partial}{\partial\theta}\ln L(x|\hat{\theta}) = 0, \text{ and}$$

$$g(1) = g(0) + \int_0^1 g'(t)\,dt,$$

whence

$$u(x|\theta_0) = g(0) = -\int_0^1 g'(t)\,dt.$$

Now, putting

$$w(x|\theta) = \frac{\partial^2}{\partial\theta^2}\ln L(x|\theta) = \sum_1^n F(x_k|\theta),$$

$$A = \frac{1}{n}\int_0^1 w(x|\theta_t)\,dt,$$

and taking into account that

$$g'(t) = w(x \mid \theta_t)\cdot(\hat{\theta} - \theta_0),$$

we have

$$u(x \mid \theta_0) = -nA(\hat{\theta} - \theta_0),$$

or

$$\frac{1}{\sqrt{n}}\,u(x|\theta_0) = -A\cdot\sqrt{n}\,(\hat{\theta} - \theta_0). \tag{1}$$

Note that the left-hand side is of the form

$$\frac{1}{\sqrt{n}}\,u(x|\theta_0) = \frac{1}{\sqrt{n}}\sum_1^n \frac{\partial}{\partial\theta}\ln f(x_k|\theta_0) = \frac{1}{\sqrt{n}}\sum_1^n (\xi_k - E_0\xi_k),$$

where

$$\xi_k = \frac{\partial}{\partial\theta}\ln f(x_k|\theta_0),\ E_0 = E_{\theta_0},\ E_0\xi_k = 0,$$

because (see A.5)

$$E_0\xi_k = \int_Y \frac{\partial}{\partial\theta}\ln f(y|\theta_0)\cdot f(y|\theta_0)\,\mu(dy)$$

$$= \int_Y \frac{\partial}{\partial\theta} f(y|\theta_0)\,\mu(dy) = \frac{\partial}{\partial\theta}\int_Y f(y|\theta_0)\,\mu(dy) = \frac{\partial}{\partial\theta} 1 = 0.$$

But, then, according to the central limit theorem of probability theory, the distribution of the RV $\frac{1}{\sqrt{n}}\, u(x \mid \theta_0)$ converges to the normal distribution $N_s(0, J_0)$ as $n \to \infty$, since the RV's $\xi_1, \xi_2, \ldots$ are i.i.d., and

$$E_0\xi_1 = 0, \quad \mathrm{var}_{\theta_0}\xi_1 = E_0\xi_1\xi_1' = J_0.$$

2°. Show that

(2) $$A \to -J_0$$

in probability as $n \to \infty$. In fact, put $A_0 = \frac{1}{n} w(x \mid \theta_0)$.

Since $A = A_0 + (A - A_0)$, it suffices to see that $A_0 \to -J_0$, and $A - A_0 \to 0$ in probability. Since

$$A_0 = \frac{1}{n}\sum_1^n \eta_k, \quad \eta_k = F(x_k|\theta_0),$$

the RV's (matrices) $\eta_1, \eta_2, \ldots$ are i.i.d., and

(3) $$E_0\eta_1 = -J_0$$

(see Item 3°), it follows from the law of large numbers that $A_0 \to -J_0$ in probability.

Further, from A6 we have

$$|A - A_0| \leqslant \left|\frac{1}{n}\int_0^1 w(x|\theta_t)\,dt - \frac{1}{n}\int_0^1 w(x|\theta_0)\,dt\right|$$

$$\leqslant \frac{1}{n}\int_0^1 |w(x|\theta_t) - w(x|\theta_0)|\,dt$$

$$\leqslant \frac{1}{n}\sum_1^n \int_0^1 |F(x_k|\theta_t) - F(x_k|\theta_0)|\,dt$$

$$\leqslant \frac{1}{n}\sum_1^n \int_0^1 c(x_k|\theta_0)\, g(\theta_t|\theta_0)\,dt = \alpha_n\beta_n,$$

where

$$\alpha_n = \frac{1}{n}\sum_1^n c(x_k|\theta_0), \quad \beta_n = \int_0^1 g(\theta_t|\theta_0)\,dt.$$

It now remains to see that

(4) $$\alpha_n \to \text{const}, \text{ and } \beta_n \to 0$$

in probability, which gives $\alpha_n\beta_n \to 0$ in probability. The latter fact implies (2).

Indeed, according to the law of large numbers, $\alpha_n \to$ const in probability, since the RV's $c(x_1 \mid \theta_0)$, $c(x_2 \mid \theta_0)$, . . . are i.i.d., and have a finite first moment.

Let us verify that $\beta_n \to 0$ in probability.

We first show that β_n is continuous as a function of points $\hat{\theta} \in \Theta$. This follows from the fact that $\forall \varepsilon > 0\ \exists\ \delta > 0$ such that $0 \leqslant g(\theta \mid \theta_0) < \varepsilon$ if $|\theta - \theta_0| < \delta$. In particular, $g(\theta_t \mid \theta_0) < \varepsilon$ when $|\hat{\theta} - \theta_0| < \delta$; therefore, $0 \leqslant \beta_n(\theta) < \varepsilon$ when $|\hat{\theta} - \theta_0| < \delta$. Now,

$$P_{\theta_0}(\beta_n \geqslant \varepsilon) \leqslant P_{\theta_0}(|\hat{\theta} - \theta_0| \geqslant \delta) \to 0 \text{ as } n \to \infty.$$

3°. We now prove (3). From A5, we have

$$0 = \frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\int_Y f(y|\theta)\,\mu(dy)$$

$$= \int \frac{\partial^2}{\partial\theta_i\,\partial\theta_j} f\,d\mu = \int \frac{\partial}{\partial\theta_i}\left[\frac{\partial \ln f}{\partial\theta_j} f\right] d\mu$$

$$= \int \left[\frac{\partial^2 \ln f}{\partial\theta_i\,\partial\theta_j} + \frac{\partial \ln f}{\partial\theta_i}\cdot\frac{\partial \ln f}{\partial\theta_j}\right] f\,d\mu$$

$$= E_\theta \frac{\partial^2 \ln f}{\partial\theta_i\,\partial\theta_j} + E_\theta \frac{\partial \ln f}{\partial\theta_i}\cdot\frac{\partial \ln f}{\partial\theta_j}$$

for $\theta = \theta_0$, whence

$$E_0\eta_1 = E_0 \frac{\partial^2}{\partial\theta^2} \ln f(x_1|\theta_0) = -E_0\left(\frac{\partial \ln f}{\partial\theta}\right)\left(\frac{\partial \ln f}{\partial\theta}\right)' = -J_0.$$

4°. Let us make use of the following simple statements.

S1. If a sequence of random matrices A_n of the same order converges in probability to a unit matrix I (i.e., $P\{\| A_n - I \| \geqslant \varepsilon\} \to 0$ for any $\varepsilon > 0$), then $A_n^{-1} \to I$ in probability (for those n, and elementary events for which the matrix A_n is non-invertible, we can assume that $A_n^{-1} = C$, where C is any fixed matrix).

S2. If a sequence of random matrices A_n converges in probability to a constant matrix A, and a sequence of random vectors ζ_n is asymptotically normal according to $N(0, B)$, then the sequence of random vectors $A_n\zeta_n$ is asymptotically normal according to $N(0, ABA')$.

It now follows from S1, (2) and the invertibility of the matrix J_0 that $-A_0^{-1} \to_0^- J^1$ in probability, but then it follows from S2 and (1) that distribution of the RV

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -A^{-1}\cdot\frac{1}{\sqrt{n}} u(x|\theta_0)$$

tends to the normal distribution $N_s(0, \Sigma)$, where

$$\Sigma = J_0^{-1}\cdot J_0\cdot (J_0^{-1})' = J_0^{-1}. \qquad \text{—Q.E.D.}$$

Remark. Instead of A3 and A4, it suffices to assume that there exists a sequence of estimators $\hat{\theta}_n$ such that

$$\frac{\partial}{\partial\theta}\ln L_n(x|\theta)=0 \text{ when } \theta=\hat{\theta}_n,$$

and

$$\hat{\theta}_n\to\theta_0 \text{ in probability } (P_{\theta_0}) \text{ as } n\to\infty.$$

Chapter 12

Point Estimation: Least Squares Method

§ 1. Examples

Example 1. Assume that observations $x_1, \ldots, x_n$ are representable in the form

$$x_i=\theta+\varepsilon_i, \qquad i=1, \ldots, n,$$

where θ is an unknown parameter, and ε_i the observation (measuring) random error.

It is required to estimate the (true) value of the parameter θ. We shall proceed from the fact that the value of $\sum_1^n \varepsilon_i^2$ should be minimum, which leads to the estimator

$$\hat{\theta}=\frac{1}{n}(x_1+\ldots+x_n)=\bar{x}.$$

Such an estimation method is called the **least squares method** (LSM).

Example 2. We now assume that observations $x_1, \ldots, x_n$ are representable in the form

$$x_i=\theta_0+\theta_1 t_i+\varepsilon_i,$$

where $\theta=(\theta_0, \theta_1)$ is an unknown parameter, the numbers $t_1, \ldots, t_n$ known, and ε_i the random observation error (e.g., the function $x(t \mid \theta)=\theta_0+\theta_1 t$ depending on θ_0 and θ_1 linearly is measured at the points $t_1, \ldots, t_n$ with measuring errors $\varepsilon_1, \ldots, \varepsilon_n$). According to the LSM, the estimator $\hat{\theta}$ minimizes the sum of the error squares

$$\sum_1^n \varepsilon_i^2=\sum_1^n (x_i-\theta_0-\theta_1 t_i)^2$$

with respect to θ.

§ 2. Normal Equation

We shall proceed from the following linear observation model, viz.,

$$x=A\theta+\varepsilon, \qquad E\varepsilon=0, \quad \operatorname{var}\varepsilon=\sigma^2\cdot I,$$

$x=(x_1, \ldots, x_n)'$ is the observation (random) vector, $\theta=(\theta_1, \ldots, \theta_s)'$ unknown parameter, $\varepsilon=(\varepsilon_1, \ldots, \varepsilon_n)'$ observation error (random) vector, and $\operatorname{var}\varepsilon=E(\varepsilon-E\varepsilon)(\varepsilon-E\varepsilon)'=E\varepsilon\varepsilon'=\{E\varepsilon_i\varepsilon_j\}$. E.g., if the RV's $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d., and $E\varepsilon_i=0$, $E\varepsilon_i^2=\sigma^2$, then $\operatorname{var}\varepsilon=\sigma^2 I$, where $I=I_n$ is the unit matrix of order $n\times n$. The matrix A of order $n\times s$ (with n rows and s columns) is regarded as known.

The LSM consists in finding $\hat{\theta}=\hat{\theta}(x)$ from the condition

$$\min_\theta(\varepsilon, \varepsilon)=\min_\theta|x-A\theta|^2=|x-A\hat{\theta}|^2.$$

Such an estimator $\hat{\theta}$ is called an **LSM-estimator.**

Theorem 1. $A'A\hat{\theta}=A'x$.
This equation is said to be **normal.**

Proof. Let $\hat{\theta}$ satisfy the normal equation. Put $u=x-A\hat{\theta}$, $v=A(\hat{\theta}-\theta)$. Then $A'u=A'x-A'A\hat{\theta}=0$, and

$$|x-A\theta|^2=|x-A\hat{\theta}+A(\hat{\theta}-\theta)|^2=|u+v|^2$$
$$=|u|^2+|v|^2+2(u, v)=|u|^2+|v|^2\geqslant|u|^2=|x-A\hat{\theta}|^2,$$

since $(u, v)=(u, A(\hat{\theta}-\theta))=(A'u, \hat{\theta}-\theta)=(0, \hat{\theta}-\theta)=0$. Meanwhile, the inequality $|x-A\theta|^2\geqslant|x-A\hat{\theta}|^2$ turns into an equality only if $v=0$, i.e., if $A\theta=A\hat{\theta}$, which gives $A'A\theta=A'x$.

Assumption. $\det(A'A)\neq 0$.

Corollary. $\hat{\theta}=Bx$, where $B=(A'A)^{-1}A'$.

Theorem 2. 1° $E\hat{\theta}=\theta$, i.e., $\hat{\theta}$ is a UE of the parameter θ.

2°. $\operatorname{var}\hat{\theta}=\sigma^2(A'A)^{-1}$.

3°. $\hat{\sigma}^2=\frac{1}{n-s}|x-A\hat{\theta}|^2$ is a UE for σ^2.

Proof. 1° $E\hat{\theta}=EBx=BEx=BA\theta=\theta$, since $BA=I_s$.

2°. Since $\hat{\theta}=Bx=B(A\theta+\varepsilon)=\theta+B\varepsilon$, we obtain

$$\operatorname{var}\hat{\theta}=E(\hat{\theta}-\theta)(\hat{\theta}-\theta)'=EB\varepsilon\cdot\varepsilon'B'=B(E\varepsilon\varepsilon')B'$$
$$=\sigma^2BB'=\sigma^2(A'A)^{-1}.$$

3°. It suffices to show that $E|x-A\hat{\theta}|^2=(n-s)\sigma^2$.

We have

$$E\,|x - A\hat{\theta}|^2 = E\,|A\theta + \varepsilon - AB\,(A\theta + \varepsilon)\,|^2$$
$$= E\,|\,\varepsilon - AB\varepsilon\,|^2 = E\,|\,\varepsilon - P\varepsilon\,|^2$$
$$= E\,(\varepsilon - P\varepsilon,\ \varepsilon - P\varepsilon) = E\,(\varepsilon - P\varepsilon,\ \varepsilon)$$
$$= E\,(\varepsilon,\ \varepsilon) - E\,(\varepsilon,\ P\varepsilon) = n\sigma^2 - E\,(\varepsilon,\ P\varepsilon),$$

where $P = AB = A\,(A'A)^{-1}A'$, and $P' = P = P^2$; therefore,

$$(\varepsilon - P\varepsilon,\ P\varepsilon) = (P'\,(\varepsilon - P\varepsilon),\ \varepsilon)$$
$$= (P\varepsilon - P^2\varepsilon,\ \varepsilon) = 0.$$

It remains to show that $E\,(\varepsilon,\ P\varepsilon) = s\cdot\sigma^2$. Note that

$$\operatorname{tr}(AB) = \operatorname{tr}(BA), \qquad (x,\ y) = \operatorname{tr}(xy'),$$

where

$$x = (x_1,\ \ldots,\ x_n)', \qquad y = (y_1,\ \ldots,\ y_n)',$$

in which case

$$E\,(\varepsilon,\ P\varepsilon) = E\ \operatorname{tr}\ \varepsilon\,(P\varepsilon)' = E\ \operatorname{tr}\ (\varepsilon\varepsilon')P' = \operatorname{tr}\,(E\varepsilon\varepsilon')P$$
$$= \sigma^2\ \operatorname{tr}\ P = \sigma^2\ \operatorname{tr}\,(AB) = \sigma^2\ \operatorname{tr}\,(BA)$$
$$= \sigma^2\ \operatorname{tr}\ I_s = \sigma^2\cdot s.$$

§ 3. Gauss-Markov Theorem

Let Δ be the class of linear UE's of a parameter θ.

Theorem. $\operatorname{var}_\theta\hat{\theta} \leqslant \operatorname{var}_\theta t\ \forall\ \theta$, and $t \in \Delta$.

Remark 1. For two matrices A and B, the inequality $A \geqslant B$ means that $A - B \geqslant 0$, i.e., the matrix $A - B$ is positive semi-definite. Note that the matrix A is positive semi-definite if and only if $(Ax,\ x) \geqslant 0$ for all x.

Proof 1°. An estimator $t\,(x)$ belongs to Δ if and only if it is of the form $t\,(x) = Dx$, where D is a certain matrix (of order $s \times n$), and

$$Et\,(x) = EDx = DEx = DA\theta = \theta\ \forall\ \theta, \text{ i.e.,}$$
$$DA = I_s$$

2°. Let $t \in \Delta$. Then

$$t\,(x) = Dx = D\,(A\theta + \varepsilon) = \theta + D\varepsilon,$$

whence

$$\operatorname{var}_\theta t = E_\theta\,(Dx - \theta)\,(Dx - \theta)' = ED\varepsilon\varepsilon'D' = D\,(E\varepsilon\varepsilon')D' = \sigma^2 DD'$$

Similarly,

$$\operatorname{var}_\theta\hat{\theta} = \sigma^2 BB' = \sigma^2\,(A'A)^{-1}.$$

It suffices to show that

$$DD' \geqslant (A'A)^{-1}$$

(making use only of the equality $DA = I_s$). Since

$$(A'A)^{-1} = DA\,(A'A)^{-1}(DA)' = DA\,(A'A)^{-1}A'D' = DPD',$$

we have $DD' - (A'A)^{-1} = D\,(I - P)\,D'$; therefore, it suffices to show that

$$Q = I_n - P \geqslant 0.$$

In fact, we obtain from the equalities $Q' = Q = Q^2$ that

$$(Qx,\ x) = (Q^2x,\ x) = (Q'Qx,\ x) = (Qx,\ Qx) \geqslant 0\ \forall\ x,$$

i.e., the matrix Q is positive semi-definite.

Remark 2. The vectors $\hat{\theta}$ and $x - A\hat{\theta}$ are uncorrelated. Note at first that the uncorrelatedness of the random vectors

$$\xi = (\xi_1,\ \ldots,\ \xi_s)' \quad \text{and} \quad \eta = (\eta_1,\ \ldots,\ \eta_n)'$$

means that of their components, i.e., $\operatorname{cov}\,(\xi_i,\ \eta_j) = 0\ \forall\ i,\ j$, $E\,(\xi - E\xi)\,(\eta - E\eta)' = 0$.

Since

$$x - A\hat{\theta} = A\theta + \varepsilon - AB\,(A\theta + \varepsilon) = \varepsilon - AB\varepsilon = \varepsilon - P\varepsilon = Q\varepsilon,$$

and

$$E\,(x - A\hat{\theta}) = QE\varepsilon = 0,$$

it suffices to show that

$$E\hat{\theta}\,(x - A\hat{\theta})' = 0.$$

We have

$$E\hat{\theta}\,(x - A\hat{\theta})' = EB\,(A\theta + \varepsilon)\,\varepsilon'Q' = EB\varepsilon\varepsilon'Q$$
$$= B\,(E\varepsilon\varepsilon')Q = \sigma^2 BQ,$$

and

$$BQ = B - BAB = B - B = 0.$$

Remark 3. Let $\lambda_1,\ \ldots,\ \lambda_s$ be the eigenvalues of a matrix $A'A$ (of order $s \times s$). As $n \to \infty$,

$$E\,|\,\hat{\theta} - \theta\,|^2 \to 0 \Leftrightarrow \min\,(\lambda_1,\ \ldots,\ \lambda_s) \to +\infty.$$

In fact,

$$E|\hat{\theta} - \theta|^2 = E\,(B\varepsilon,\ B\varepsilon) = E\ \operatorname{tr}\ B\varepsilon\,(B\varepsilon)'$$
$$= \operatorname{tr}\,B\,(E\varepsilon\varepsilon')\,B' = \sigma^2\ \operatorname{tr}\ BB'$$
$$= \sigma^2\ \operatorname{tr}\,(A'A)^{-1}A'\cdot A\,(A'A)^{-1}$$
$$= \sigma^2\ \operatorname{tr}\,(A'A)^{-1}$$
$$= \sigma^2\left[\frac{1}{\lambda_1} + \ldots + \frac{1}{\lambda_s}\right].$$

Note, in addition, that, according to the Chebyshev inequality, for the consistency of the LSM-estimator $\hat{\theta}$, it suffices that we should have $E\,|\,\hat{\theta}-\theta\,|^2\to 0$.

§ 4. Linear Regression Model

Assume that observations $x_1, \ldots, x_n$ are representable in the form

$$x_i = f(t_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where the function f is known up to parameter $\theta = (\theta_1, \ldots, \theta_s)'$ on which it depends linearly, the numbers $t_1, \ldots, t_n$ (at which the value of f is measured) are known, observational (measuring) errors $\varepsilon_1, \ldots, \varepsilon_n$ are pairwise uncorrelated, and $E\varepsilon_i = 0$, $\operatorname{var}\varepsilon_i = \sigma^2$.

It is required to estimate the function $f(t) = f(t, \theta)$ and, in particular, $f(t, \theta)$ for a given number t.

Such a problem is given rise if we would like to determine the relation between t and x.

As the function $f(t, \theta)$, the following functions are often taken, viz.,

$$f(t, \theta) = \theta_0 + \theta_1 t,$$
$$f(t, \theta) = \theta_0 + \theta_1 t + \ldots + \theta_m t^m,$$
$$f(t, \theta) = \theta_0 + \theta_1 \ln t,$$
$$f(t, \theta) = \theta_0 + \theta_1 t + \theta_2 \frac{1}{t}.$$

The model $g(x_i) = f(t_i, \theta) + \varepsilon_i$, $i = 1, \ldots, n$, is sometimes employed (it is more general than (1)) where g is a bijection.

To estimate the parameter θ, we can make use of the LSM, and if $\hat{\theta}$ is an LSM-estimator, then the function (of t)

$$x = f(t, \hat{\theta})$$

is said to be the **statistical regression function** (or **line**, or **curve**).

Example 1 (Linear Regression). Assume that

$$g(x) = x, \quad f(t, \theta) = \theta_0 + \theta_1 t,$$

in which case

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}.$$

The normal equation $(A'A)\hat{\theta} = A'\tilde{x}$, $\tilde{x} = (x_1, \ldots, x_n)'$ takes the form

$$\begin{bmatrix} n & \sum t_i \\ \sum t_i & \sum t_i^2 \end{bmatrix} \begin{bmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{bmatrix} = \begin{bmatrix} \sum x_i \\ \sum t_i x_i \end{bmatrix},$$

whence

$$\hat{\theta}_0 = \frac{\frac{1}{n}\sum t_i^2 \cdot \frac{1}{n}\sum x_i - \frac{1}{n}\sum t_i \cdot \frac{1}{n}\sum t_i x_i}{\frac{1}{n}\sum t_i^2 - \left(\frac{1}{n}\sum t_i\right)^2} = \frac{1}{n}\sum x_i - \hat{\theta}_1 \cdot \frac{1}{n}\sum t_i,$$

$$\hat{\theta}_1 = \frac{\frac{1}{n}\sum t_i x_i - \frac{1}{n}\sum t_i \cdot \frac{1}{n}\sum x_i}{\frac{1}{n}\sum t_i^2 - \left(\frac{1}{n}\sum t_i\right)^2}.$$

The statistical regression line is

$$x = \hat{\theta}_0 + \hat{\theta}_1 t.$$

Example 2 (Polynomial Regression). Assume that

$$g(x) = x, \quad f(t, \theta) = \theta_0 + \theta_1 t + \ldots + \theta_m t^m,$$

in which case

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_m \end{bmatrix}, \quad A = \begin{bmatrix} 1 & t_1 & t_1^2 & \vdots & t_1^m \\ 1 & t_2 & t_2^2 & \vdots & t_2^m \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & t_n & t_n^2 & \vdots & t_n^m \end{bmatrix}.$$

The normal equation takes the form

$$\begin{bmatrix} n & \sum t_i & \sum t_i^2 & \vdots & \sum t_i^m \\ \sum t_i & \sum t_i^2 & \sum t_i^3 & \vdots & \sum t_i^{m+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum t_i^m & \sum t_i^{m+1} & \sum t_i^{m+2} & \vdots & \sum t_i^{2m} \end{bmatrix} \begin{bmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_m \end{bmatrix} = \begin{bmatrix} \sum x_i \\ \sum t_i x_i \\ \vdots \\ \sum t_i^m x_i \end{bmatrix},$$

and the statistical regression curve is

$$x = \hat{\theta}_0 + \hat{\theta}_1 t + \ldots + \hat{\theta}_m t^m.$$

Example 3 (Exponential Regression). Assume now that

$$g(x) = \ln x, \quad f(t, \theta) = \theta_0 + \theta_1 t,$$

in which case the normal equation is of the form

$$\begin{bmatrix} n & \sum t_i \\ \sum t_i & \sum t_i^2 \end{bmatrix} \begin{bmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{bmatrix} = \begin{bmatrix} \sum \ln x_i \\ \sum t_i \ln x_i \end{bmatrix},$$

whence

$$\hat{\theta}_0 = \frac{\frac{1}{n}\sum t_i^2 \cdot \frac{1}{n}\sum \ln x_i - \frac{1}{n}\sum t_i \cdot \frac{1}{n}\sum t_i \ln x_i}{\frac{1}{n}\sum t_i^2 - \left(\frac{1}{n}\sum t_i\right)^2}$$

$$= \frac{1}{n}\sum \ln x_i - \hat{\theta}_1 \frac{1}{n}\sum t_i,$$

$$\hat{\theta}_1 = \frac{\frac{1}{n}\sum t_i \ln x_i - \frac{1}{n}\sum t_i \cdot \frac{1}{n}\sum \ln x_i}{\frac{1}{n}\sum t_i^2 - \left(\frac{1}{n}\sum t_i\right)^2}.$$

The statistical regression curve is

$$\ln x = \hat{\theta}_0 + \hat{\theta}_1 t, \text{ or } x = \exp\{\hat{\theta}_0 + \hat{\theta}_1 t\}.$$

Remark. 1°. Assume that two RV's x and t are involved in the relation $x = \theta_0 + \theta_1 t$, and take values from $\{x_1, \ldots, x_n\}$ and $\{t_1, \ldots, t_n\}$, respectively; meanwhile,

$$P(t = t_k,\ x = x_k) = \frac{1}{n},$$

$$x_k = \theta_0 + \theta_1 t_k,\ k = 1, \ldots, n.$$

Then

$$Ex = \theta_0 + \theta_1 Et,\ \operatorname{cov}(x, t) = \theta_1 \operatorname{var} t,$$

whence

$$\theta_0 = Ex - \theta_1 Et = \frac{1}{n}\sum x_i - \theta_1 \frac{1}{n}\sum t_i,$$

$$\theta_1 = \frac{\operatorname{cov}(x, t)}{\operatorname{var} t} = \frac{Ext - ExEt}{Et^2 - (Et)^2}$$

$$= \frac{\frac{1}{n}\sum x_i t_i - \frac{1}{n}\sum x_i \cdot \frac{1}{n}\sum t_i}{\frac{1}{n}\sum t_i^2 - \left(\frac{1}{n}\sum t_i\right)^2}.$$

We see that $\theta = (\theta_0, \theta_1)'$ coincides with the LSM-estimator $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1)'$ of Example 1, which permits us to remember the LSM-estimator form easier.

2°. The same technique is also suitable in the general case. In fact, assume now that two RV's x and t are involved in the relation

$$x = \theta_1 e_1(t) + \ldots + \theta_s e_s(t),$$

and

$$P(t = t_k,\ x = x_k) = \frac{1}{n},$$

$$x_k = \theta_1 e_1(t_k) + \ldots + \theta_s e_s(t_k),\ k = 1, \ldots, n.$$

Then

$$\begin{cases} Exe_1(t) = \theta_1 Ee_1(t)\, e_1(t) + \ldots + \theta_s Es_s(t)\, e_1(t) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ Exe_s(t) = \theta_1 Ee_1(t)\, e_s(t) + \ldots + \theta_s Ee_s(t)\, e_s(t) \end{cases}$$

where

$$Exe_i(t) = \frac{1}{n}\sum_{k=1}^{n} x_k e_i(t_k),\ Ee_i(t)\, e_j(t) = \frac{1}{n}\sum_{1}^{n} e_i(t_k)\, e_j(t_k).$$

This system of linear equations in $\theta = (\theta_1, \ldots, \theta_s)'$ coincides (show!) with the normal equation of the model

$$x_k = \theta_1 e_1(t_k) + \ldots + \theta_s e_s(t_k) + \varepsilon_k,\ k = 1, \ldots, n,$$

$$E\varepsilon_k = 0,\ E\varepsilon_i \varepsilon_j = \sigma^2 \delta_{ij}.$$

§ 5. Normal Error Distribution

In this section, we assume that the error vector $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ has the normal distribution. We show that

1°. $\frac{1}{\sigma^2}|A(\theta - \hat{\theta})|^2$ has the $\chi^2(s)$-distribution,

2°. $\frac{1}{\sigma^2}|x - A\hat{\theta}|^2$ has the $\chi^2(n-s)$-distribution,

3°. $\dfrac{\frac{1}{s}|A(\theta - \hat{\theta})|^2}{\frac{1}{n-s}|x - A\hat{\theta}|^2}$ has the Snedecor-Fisher distribution $F_{s,\, n-s}$,

and

4°. the LSM-estimator coincides with the MLE.

We make use of the following statements, assuming that an RV ξ has the normal distribution $N(0, e)$ in the n-dimensional space R^n.

S1. Let H be a linear subspace of the space R^n. Then the RV $|\operatorname{pr}_H \xi|^2$ has the distribution $\chi^2(\dim H)$, where $\operatorname{pr}_H \xi$ is the projection of ξ onto H, and $|x|$ the length of a vector $x \in R^n$.

S2. Let H_1 and H_2 be two orthogonal linear subspaces of the space R^n. Then the RV's

$$\operatorname{pr}_{H_1}\xi \quad \text{and} \quad \operatorname{pr}_{H_2}\xi$$

are independent.

S3. With the data of the previous statement, the RV

$$\frac{\frac{1}{\dim H_1}|\operatorname{pr}_{H_1}\xi|^2}{\frac{1}{\dim H_2}|\operatorname{pr}_{H_2}\xi|^2}$$

has the Shedecor-Fisher distribution $F_{\dim H_1,\, \dim H_2}$.

Proof of S1. If U is an orthogonal transformation of the space R^n, then the RV $U\xi$ has the distribution $N(0, UIU') = N(0, I)$. In particular, we select an orthonormal basis $\{e_1, \ldots, e_n\}$ for R^n, so that $\{e_1, \ldots, e_k\}$ may be a basis for H, $k = \dim H$. Put $x_i = (\xi, e_i)$, $x = (x_1, \ldots, x_n)'$. The transformation $\xi \to x$ is orthogonal; therefore, the random vector x has the normal distribution $N(0, I)$. And since $\mathrm{pr}_H \xi = x_1 e_1 + \ldots + x_k e_k$, the RV's $x_1, \ldots, x_k$ are independent and have the normal distribution $N(0, 1)$, the RV $|\mathrm{pr}_H \xi|^2 = x_1^2 + \ldots + x_k^2$ has the $\chi^2(k)$-distribution.

Proof of S2. We select an orthonormal basis $\{e_1, \ldots, e_n\}$ so that $\{e_1, \ldots, e_{i_1}\}$ and $\{e_{i_1+1}, \ldots, e_{i_1+i_2}\}$ may be bases for H_1 and H_2, respectively, with $i_1 = \dim H_1$, $i_2 = \dim H_2$. Then

$$\eta_1 = \mathrm{pr}_{H_1}\xi = \sum_{i=1}^{i_1} x_i e_i,$$

$$\eta_2 = \mathrm{pr}_{H_2}\xi = \sum_{i=i_1+1}^{i_1+i_2} x_i e_i, \quad x_i = (\xi, e_i).$$

The random vector $x = (x_1, \ldots, x_n)'$ obtained from ξ by means of a certain orthogonal transformation has the normal distribution $N(0, I)$; therefore, the RV's $x_1, \ldots, x_n$ are independent, which implies the independence of the random vectors η_1 and η_2.

Proof of S3. S3 immediately follows from S1, S2 and the definition of the Snedecor-Fisher distribution.

Proof of 1°, 2° and 3°. With the notation of §§ 2 and 3, we have

$$A(\hat{\theta} - \theta) = P\varepsilon, \qquad x - A\hat{\theta} = Q\varepsilon,$$

where $P : R^n \to R^n$ and $Q : R^n \to R^n$ are symmetric projection operators (i.e., $P^2 = P$, $Q^2 = Q$), and $P + Q = I$ is the identity operator. Meanwhile, the rank of the matrix P equals s (as the rank of the product $A \cdot (A'A)^{-1} \cdot A'$ of matrices of orders $n \times s$, $s \times s$ and $s \times n$, respectively, and of rank s). Put

$$H_1 = \{Px : x \in R^n\}, \qquad H_2 = \{Qx : x \in R^n\},$$

where H_1 and H_2 are linear subspaces of the space R^n. Meanwhile, $H_1 \perp H_2$. In fact,

$$(Px, Qy) = (x, P'Qy) = (x, PQy) = 0,$$

since $PQ = P - P^2 = 0$. Because $P + Q = I$, we have $R^n = H_1 + H_2$; therefore, $\dim H_2 = n - \dim H_1 = n - s$.

Put $\xi = \frac{1}{\sigma}\varepsilon$. Then the random vector ξ has the normal distribution $N(0, I)$. It remains to make use of S1 and S3, while taking into account that

$$|\mathrm{pr}_{H_1}\xi|^2 = \frac{1}{\sigma^2}|P\varepsilon|^2 = \frac{1}{\sigma^2}|A(\hat{\theta} - \theta)|^2,$$

$$|\mathrm{pr}_{H_2}\xi|^2 = \frac{1}{\sigma^2}|Q\varepsilon|^2 = \frac{1}{\sigma^2}|x - A\hat{\theta}|^2.$$

Proof of 4°. The LF of the random vector $x = A\theta + \varepsilon$ has the form

$$L(x \mid \theta) = (2\pi)^{-\frac{n}{2}} \cdot \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}(x - A\theta, x - A\theta)\right\};$$

therefore, the MLE minimizes

$$(x - A\theta, x - A\theta) = |x - A\theta|^2$$

with respect to θ. But the LSM-estimator does the same.

Remark. Since

$$L(x \mid \theta) = (2\pi)^{-\frac{n}{2}}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}(|x - A\hat{\theta}|^2 + |A(\hat{\theta} - \theta)|^2)\right\}$$

$$= (2\pi)^{-\frac{n}{2}}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}(|x - A\hat{\theta}|^2 + |A\hat{\theta}|^2) + \frac{1}{\sigma^2}(\hat{\theta}, A'A\theta)\right\} \exp\left\{-\frac{1}{2\sigma^2}|A\theta|^2\right\},$$

and the matrix $A'A$ (of order $s \times s$) is invertible, the statistic

$$T(x) = (\hat{\theta}, \hat{\sigma}^2), \text{ where } \hat{\sigma}^2 = \frac{1}{n-s}|x - A\hat{\theta}|^2$$

is sufficient and complete. Therefore (see Theorem 2, § 2), $\hat{\sigma}^2$ is the MVUE for σ^2, and $\hat{\theta}$ that of the parameter θ.

Chapter 13

Interval Estimation

§ 1. Definition

Let $(X, \mathcal{A}, \mathcal{P} = \{P_\theta, \theta \in \Theta\})$ be a statistical structure again. A mapping S

$$X \ni x \to S(x) \subset \Theta$$

such that

(1) $$S^{-1}(\theta) = \{x : \theta \in S(x)\} \in \mathcal{A}$$

for all $\theta \in \Theta$ is called an **interval estimator.**

Remark 1. Condition (1) is required for probabilities

(2) $$P_\theta\{\theta \in S(x)\} \; \forall\, \theta \in \Theta$$

to be defined (conversely, from the fact that probabilities (2) are defined follows (1)). In addition, (2) is the probability that the random

set $S(x)$ contains the true value of the parameter θ (as always, the observation (RV) x has a distribution from the family $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, and the expression "$\theta$ is the true value of a parameter" means that "the observation x has the distribution P_θ").

Remark 2. The mapping $S(x)$, is also called a **confidence estimator**, or **multiple estimator**, or **confidence set**.

The number

$$1 - \alpha = \inf_\theta P_\theta \{\theta \in S(x)\}$$

is called the **confidence coefficient** of an interval estimator $S(x)$, and α the (significance) **level** of the interval estimator. In particular,

$$P_\theta \{\theta \in S(x)\} \geqslant 1 - \alpha \; \forall \, \theta \in \Theta.$$

An interval estimator of a level α is said to be **free** if

$$P_\theta \{\theta \in S(x)\} = 1 - \alpha \; \forall \, \theta.$$

Example 1. Assume that RV's $x_1, \ldots, x_n$ are i.i.d. according to $N(\theta, \delta^2)$ with known variance $\sigma^2 \neq 0$. Since the RV

$$\sqrt{n}\,\frac{\bar{x}-\theta}{\sigma},$$

where $\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n)$, has the normal distribution $N(0, 1)$, we obtain

$$P_\theta\left\{|\theta - \bar{x}| \leqslant c_\alpha \frac{\sigma}{\sqrt{n}}\right\} = P_\theta\left\{\left|\sqrt{n}\,\frac{\bar{x}-\theta}{\sigma}\right| \leqslant c_\alpha\right\}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-c_\alpha}^{c_\alpha} e^{-\frac{u^2}{2}}\, du = 1 - \alpha.$$

In particular, if $c_\alpha = 3$, then $\alpha = 0.00270\ldots$, and

$$P_\theta\{|\theta - x| \leqslant 3\sigma/\sqrt{n}\} = 0.99730\ldots \; \forall \theta$$

(Three-Sigma Rule), in which case

$$S(x) = \left[\bar{x} - c_\alpha \frac{\sigma}{\sqrt{n}},\ \bar{x} + c_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

is a free interval estimator of the level α, and the probability that the interval $S(x)$ contains the true value of the parameter θ equals $1-\alpha$.

Example 2 (*continued*). Assume, in addition, that the variance of σ^2 is also unknown. Put

$$s^2 = \frac{1}{n-1} \sum_1^n (x_k - \bar{x})^2, \quad s^2 > 0 \text{ a.c.}$$

Since the RV's

$$\sqrt{n}\,\frac{\bar{x}-\theta}{\sigma} \quad \text{and} \quad (n-1)\frac{s^2}{\sigma^2}$$

are independent with the normal $N(0, 1)$ and $\chi^2(n-1)$-distribution, respectively, the RV

$$\sqrt{n}\,\frac{\bar{x}-\theta}{s}$$

has the Student distribution $S(n-1)$ with $n-1$ degrees of freedom. Therefore,

$$P_{(\theta,\sigma)}\left\{|\theta - \bar{x}| \leqslant c_\alpha \frac{s}{\sqrt{n}}\right\} = P_{(\theta,\sigma)}\left\{\left|\sqrt{n}\,\frac{\bar{x}-\theta}{s}\right| \leqslant c_\alpha\right\}$$

$$= F_{n-1}(c_\alpha) - F_{n-1}(-c_\alpha) = 1 - \alpha, \quad F_{n-1}(c_\alpha) = 1 - \frac{\alpha}{2},$$

where F_{n-1} is the Student distribution with $n-1$ degrees of freedom. For $c_\alpha = 3$, we have

$$P_{(\theta,\sigma)}\left\{|\theta - \bar{x}| \leqslant \frac{3s}{\sqrt{n}}\right\} > 0.99$$

when $n \geqslant 15$ (and the left-hand side does not depend on (θ, σ) for a fixed n).

§ 2. Interval Maximum Likelihood Estimator

Let $L(x \mid \theta)$ be an LF. An interval estimator $S(x)$ is called an **interval maximum likelihood estimator** if $\theta_1 \in S(x)$, $L(x \mid \theta_2) \geqslant L(x \mid \theta_1)$ implies $\theta_2 \in S(x)$.

Usually, an interval MLE has the form

(1) $\quad S(x) = \{\theta : L(x \mid \theta) \geqslant L(x)\}$,

where $L(x)$ is a certain function of x. An estimator of form (1) is at least an interval MLE (it should also be assumed that $S^{-1}(\theta) \in \mathcal{A}$ for all θ, provided that, for any θ, the functions $L(x \mid \theta)$ and $L(x)$ should be Borel). Conversely, if $S(x)$ is an interval MLE, and

$$L(x) = \min_{\theta \in S(x)} L(x \mid \theta)$$

(for any x, the lower bound of $L(x \mid \theta)$ is attained with respect to $\theta \in S(x)$!), then $S(x)$ has form (1). If $\hat{\theta}(x)$ is an MLE, and $S(x) \neq \varnothing$, then $\hat{\theta}(x) \in S(x)$ (see Fig. 15).

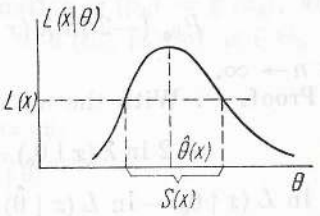

Fig. 15

§ 3. Likelihood Ratio Principle

A useful and frequently applied method of obtaining interval MLE's was suggested by Neyman and Pearson in the form of the **likelihood ratio principle** which states that an interval MLE of a level α has the form

$$S(x) = \{\theta : \lambda(x \mid \theta) \geqslant c_\alpha\},$$

where

$$\lambda(x \mid \theta) = L(x \mid \theta)/\sup_{\omega \in \Theta} L(x \mid \omega).$$

A complicated stage in applying this principle is the computation of c_α for each level α. However, if $x = (x_1, \ldots, x_n)$ and the observations $x_1, \ldots, x_n$ are i.i.d., then, for sufficiently general assumptions and each θ, the distribution of the RV $-2\ln\lambda(x \mid \theta)$ tends to the χ^2-distribution (not depending on θ) as $n \to \infty$. And since the χ^2-distribution is tabulated, for large n, the number c_α can be obtained from the relation

$$\begin{aligned} P_\theta\{\theta \in S(x)\} &= P_\theta\{\lambda(x \mid \theta) \geqslant c_\alpha\} \\ &= P_\theta\{-2\ln\lambda(x \mid \theta) \leqslant -2\ln c_\alpha\} \\ &\simeq P\{\chi^2 \leqslant -2\ln c_\alpha\} = 1-\alpha. \end{aligned}$$

As a rule, the number of degrees of freedom in the χ^2-distribution is determined simply (see the next section).

§ 4. Likelihood Ratio Principle Asymptotic Property

Theorem. Let A1-A7, § 4, Ch. 11, be valid. Then, for any number,

$$P_{\theta_0}\{-2\ln\lambda(x \mid \theta_0) < t\} \to P \Leftrightarrow \{\chi^2(s) < t\}$$

as $n \to \infty$.

Proof. 1°. With the notation of § 4, Ch. 11, we have

$$\begin{aligned} -2\ln\lambda(x \mid \theta_0) &= 2[\ln L(x \mid \hat\theta) - \ln L(x \mid \theta_0)], \\ \ln L(x \mid \theta_0) - \ln L(x \mid \hat\theta) &= \left(\frac{\partial}{\partial\theta}\ln L(x \mid \hat\theta),\ \theta_0 - \hat\theta\right) \\ &\quad + \frac{1}{2!}(nA_0(\theta_0 - \hat\theta),\ \theta_0 - \hat\theta) + R \\ &= \frac{1}{2}(A_0\sqrt{n}(\hat\theta - \theta_0),\ \sqrt{n}(\hat\theta - \theta_0)) + R, \end{aligned}$$

where $A_0 \to -J_0$, $R \to 0$ in probability $P_0 = P_{\theta_0}$ as $n \to \infty$.

2°. The RV $\sqrt{n}(\hat\theta - \theta_0)$ converges weakly (in distribution) to an RV ξ_0 having the normal distribution $N(0, J_0^{-1})$. Therefore, the RV $-2\ln\lambda(x \mid \theta_0)$ converges weakly to the RV $(J_0\xi_0, \xi_0)$.

3°. It remains to show that the RV $(J_0\xi_0, \xi_0)$ has the distribution $\chi^2(s)$. In fact, since the matrix J_0 is positive definite, it is representable in the form $J_0 = B^2$, where $B' = B$. If $\eta_0 = B\xi_0$, then

$$(J_0\xi_0, \xi_0) = (\eta_0, \eta_0),$$

and the RV η_0 has the distribution $N_s(0, I_s)$, where I_s is the unit matrix of order $s \times s$ (see Property F, § 16, Ch. 2).

§ 5. Generalized Likelihood Ratio Principle

In §§ 3, 4, we have suggested a method for obtaining an interval (maximum likelihood) estimator of a parameter θ. Now, we indicate a technique for finding an interval estimator $g(\theta)$, where g is a mapping of Θ into $G \subset R^n$. In addition to A1-A7, § 4, Ch. 11, we consider the following assumption regarding g.

A8. There exist two sets Θ_0, Θ_1 and a bijection

$$\Theta_0 \times \Theta_1 \ni (\alpha, \beta) \to \theta = \theta(\alpha, \beta) \in \Theta \subset R^s$$

such that

(1) $g(\theta(\alpha, \beta)) = \bar g(\beta)$ does not depend on α, and the mapping $\bar g\colon \Theta_1 \to G$ is bijective,

(2) the set $\Theta_0 \subset R^k$ is (locally) convex, and the rank of the matrix

$$C = C(\beta_0) = \frac{\partial\theta}{\partial\alpha} = \left\{\frac{\partial\theta_j}{\partial\alpha_i}\right\}$$

equals $k = s - r$ for any $\alpha \in \Theta_0$ and $\beta = \beta_0$, where the element β_0 is (uniquely) determined by the equation $\bar g(\beta_0) = g(\theta_0)$, and

(3) when $\beta = \beta_0$, the function $\bar\theta(\alpha) = \theta(\alpha_0, \beta_0)$ of $\alpha \in \Theta_0$ is twice continuously differentiable.

For each $g \in G$, we put

$$\lambda(x \mid g) = \frac{\max\limits_{\theta:g(\theta)=g} L(x \mid \theta)}{\max\limits_{\theta} L(x \mid \theta)}.$$

Theorem. For each number t,

$$P_{\theta_0}\{-2\ln\lambda(x \mid g_0) < t\} \to P\{\chi^2(r) < t\},$$

where $g_0 = g(\theta_0)$, as $n \to \infty$.

Remark. If A1-A8 are realized for any $\theta_0 \in \Theta$, then

$$P_0\{g(\theta) \in G(x)\} \to 1 - \gamma \text{ as } n \to \infty \ \forall\theta,$$

where

$$G(x) = \{g \in G : \lambda(x \mid g) \geqslant c_\gamma\},$$

and the number c_γ is determined from the equation

$$P\{\chi^2(r) \leqslant -2\ln c_\gamma\} = 1-\gamma,\ 0<\gamma<1,$$

which means that $G(x)$ is an asymptotically interval estimator for $g(\theta)$ of the level γ. In fact,

$$P_\theta\{g(\theta)\in G(x)\} = P_\theta\{\lambda(x|g(\theta)) \geqslant c_\gamma\}$$
$$= P_\theta\{-2\ln\lambda(x|g(\theta)) \leqslant -2\ln c_\gamma\}$$
$$\to P\{\chi^2(r) \leqslant -2\ln c_\gamma\} = 1-\gamma\ \forall\theta.$$

Corollary. Assume that $\Theta_0 \subset \Theta \subset R^s$, and the set Θ_0 is convex, whereas the minimal linear subspace (of the space R^s) containing Θ_0 has dimension k. Put

$$\lambda(x) = \frac{\max\limits_{\theta\in\Theta_0} L(x|\theta)}{\max\limits_{\theta\in\Theta} L(x|\theta)}.$$

If A1-A7 are realized for all $\theta_0 \in \Theta_0$, then

$$P_\theta\{-2\ln\lambda(x) < t\} \to P\{\chi^2(s-k) < t\}$$

for each $\theta \in \Theta_0$ and each number t as $n \to \infty$.

In fact, in our case,

$$\theta = \theta(\alpha, \beta) = \alpha + \beta, \qquad \alpha \in \Theta_0 \text{ and } \beta \in \Theta_1,$$

where Θ_1 is contained in the orthogonal complement of Θ_0, and $g(\theta) = \alpha$ (it is thereby assumed that $\Theta = \Theta_0 + \Theta_1 = \{\alpha+\beta : \alpha\in\Theta_0, \beta\in\Theta_1\}$).

Proof. 1°. Proceeding from the LF $L(\theta) = L(x \mid \theta)$, we put

$$L^*(\alpha) = L^*(x \mid \alpha) = L(x \mid \theta(\alpha, \beta_0)),$$
$$f^*(y \mid \alpha) = f(y \mid \theta(\alpha, \beta_0)),\ \alpha\in\Theta_0,$$

$\hat\theta$ is the MLE for θ, $\hat\alpha$ the MLE for α, and

$$\lambda(x|\theta) = \frac{L(x|\theta)}{L(x|\hat\theta)},\ \lambda^*(x|\alpha) = \frac{L^*(x|\alpha)}{L^*(x|\hat\alpha)}.$$

Then, for the element $\alpha_0 \in \Theta_0$ determined from the equation $\theta_0 = \theta(\alpha_0, \beta_0)$, we have

$$\lambda(x|g_0) = \frac{L^*(x|\hat\alpha)}{L(x|\hat\theta)} = \frac{\lambda(x|\theta_0)}{\lambda^*(x|\alpha_0)},$$

and

$$-2\ln\lambda(x|g_0) = -2\ln\lambda(x|\theta_0) + 2\ln\lambda^*(x|\alpha_0)$$
$$= -2[\ln L(x|\theta_0) - \ln L(x|\hat\theta)]$$
$$+2[\ln L^*(x|\alpha_0) - \ln L^*(x|\hat\alpha)].$$

2°. We have

$$\ln L(\theta_0) - \ln L(\hat\theta) = \frac{1}{2!}(A_0 x, x) + R,$$

(1)
$$\ln L^*(\alpha_0) - \ln L(\hat\alpha) = \frac{1}{2!}(A_0^* y, y) + R^*,$$

where

$$x = \sqrt{n}(\hat\theta - \theta_0),\ y = \sqrt{n}(\hat\alpha - \alpha_0),$$
$$A_0 = \left\{\frac{1}{n}\cdot\frac{\partial^2 \ln L(\theta_0)}{\partial\theta_i\partial\theta_j}\right\},$$
$$A_0^* = \left\{\frac{1}{n}\cdot\frac{\partial^2 \ln L^*(\alpha_0)}{\partial\alpha_i\partial\alpha_j}\right\},$$
$$R \to 0,\ A_0 \to -J_0 \text{ in probability } P_0.$$

3°. Replacing

$$\theta, \Theta, f(y \mid \theta), L(x \mid \theta)$$

by

$$\alpha, \Theta_0, f^*(y \mid \alpha), L^*(x \mid \alpha),$$

respectively, we obtain that A1-A7, (§ 4, Ch. 11) are also realized. E.g., A7 is valid, since

$$J_0^* = J^*(\alpha_0) = \left\{E_{\alpha_0}\frac{\partial \ln f^*(x_1|\alpha)}{\partial\alpha_i}\cdot\frac{\partial \ln f^*(x_1|\alpha)}{\partial\alpha_j}\right\}_{\alpha=\alpha_0}$$
$$= E_{\alpha_0}\left(\frac{\partial \ln f^*}{\partial\alpha}\right)\left(\frac{\partial \ln f^*}{\partial\alpha}\right)',$$
$$\frac{\partial \ln f^*}{\partial\alpha} = C\cdot\frac{\partial \ln f}{\partial\theta} = \left\{\sum_m \frac{\partial \ln f}{\partial\theta_m}\frac{\partial\theta_m}{\partial\alpha_i}\right\},$$

and the rank of the matrix C equals k (therefore, the matrix J_0^* of order $k\times k$ is positive definite).

To verify A6, we should make use of the following statement, viz., if

$$w(\alpha) = W(\theta, (\alpha)),$$

then

$$\frac{\partial w}{\partial\alpha} = C\frac{\partial W}{\partial\theta},$$
$$\frac{\partial^2 w}{\partial\alpha^2} = C\frac{\partial^2 W}{\partial\theta^2}C' + \sum_m \frac{\partial W}{\partial\theta_m}C_m,$$

where

$$\frac{\partial w}{\partial\alpha} = \left\{\frac{\partial w}{\partial\alpha_i}\right\},\ \frac{\partial^2 w}{\partial\alpha^2} = \left\{\frac{\partial^2 w}{\partial\alpha_i\partial\alpha_j}\right\},$$
$$C = \left\{\frac{\partial\theta_j}{\partial\alpha_i}\right\},$$
$$\theta = (\theta_1, \ldots, \theta_s),\ C_m = \frac{\partial^2\theta_m}{\partial\alpha^2}.$$

4°. Due to Statement 3° and formula (1), we have, similarly, that

$$R^* \to 0,\ A_0^* \to J_0^* \text{ as } n \to \infty$$

in probability $P_0 = P_{\theta_0} = P^*_{\alpha_0}$.

5°. In addition, putting

$$u = \frac{1}{n}\frac{\partial \ln L(\theta_0)}{\partial \theta}, \qquad v = \frac{1}{n}\frac{\partial \ln L^*(\alpha)}{\partial \alpha},$$

we have $v = Cu$, and

$$u \overset{\text{as.}}{=} J_0 x, \qquad v \overset{\text{as.}}{=} J_0^* y,$$

where the notation $\xi_n \overset{\text{as.}}{=} \eta_n$ means that the RV's ξ_n and η_n have one and the same weak limit (in distribution). Note that the RV's u and v are asymptotically normal to $N_s(0, J_0)$ and $N_k(0, J_0^*)$, respectively.

6°. We now have

$$\begin{aligned} -2\ln\lambda(x \mid g_0) &\overset{\text{as.}}{=} (J_0 x, x) - (J_0^* y, y) \\ &\overset{\text{as.}}{=} (u, J_0^{-1} u) - (d, J^{*-1}_0 v) \\ &= (J^{-1}_0 u, v) - (C' J_0^{*-1} Cu, u) = (Au, u), \end{aligned}$$

where

$$A = J^{-1}_0 - C' J_0^{*-1} C.$$

Thus,

$$-2\ln\lambda(x \mid g_0) \overset{\text{as.}}{=} (A\xi, \xi),$$

where the RV ξ has the normal distribution $N(0, B)$, $B = J_0$. It now suffices to show that the RV $(A\xi, \xi)$ has the distribution $\chi^2(r)$.

7°. Note that

$$A = ABA.$$

In fact, since $J_0^* = CBC'$, we get

$$\begin{aligned} ABA &= (I - C' J_0^{*-1} CB)(B^{-1} - C' J_0^{*-1} C) \\ &= B^{-1} - C' J_0^{*-1} C - C' J_0^{*-1} C + C'(J_0^{*-1} CBC') J_0^{-1} C \\ &= B^{-1} - C' J_0^{*-1} C = A. \end{aligned}$$

8°. Let us make use of the following statement. If an RV ξ has the normal distribution $N(0, B)$, then the RV $(A\xi, \xi)$ has the distribution $\chi^2(m)$ if (and only if) $A = ABA$, and $\operatorname{tr}(AB) = m$ (show!).

9°. In our case,

$$\begin{aligned} \operatorname{tr}(AB) &= \operatorname{tr}(I_s - C' J_0^{*-1} CB) = s - \operatorname{tr}(J_0^{*-1} \cdot CBC') \\ &= s - \operatorname{tr} I_k = s - k = r, \end{aligned}$$

where I_k is the unit matrix of order $k \times k$.

§ 6. Technique of Likelihood Ratio Principle Application

A. Likelihood ratio principle as a method of obtaining interval estimators.

By this principle, an interval (maximum likelihood) estimator $G(x)$ for $g(\theta)$ of a level γ is determined by the equalities

$$G(x) = \{g : \lambda(x \mid g) \geqslant c_\gamma\},$$

$$\lambda(x \mid g) = \max_{\theta:\, g(\theta)=g} L(x \mid \theta) / \max_{\theta} L(x \mid \theta),$$

$$P_\theta\{g(\theta) \in G(x)\} \geqslant 1 - \gamma \quad \forall\ \theta.$$

Since, for large n, the RV $-2\ln\lambda(x \mid g(\theta))$ has (approximately) the distribution $\chi^2(r)$, where r is the dimension of the vector function $g(\theta)$ (see the theorem of § 5), the constant c_γ can be determined from the condition

$$P\{\chi^2(r) \leqslant -2\ln c_\gamma\} = 1 - \gamma.$$

Unfortunately, the problem of the rate of convergence of the distribution $-2\ln\lambda(x \mid g(\theta))$ to the corresponding χ^2-distribution is not investigated; therefore, the existing recommendations as to which n should be regarded as "large" are not fully justified. Nevertheless, this method is already employed when $n \geqslant 100$ (and even $n \geqslant 50$).

In certain cases, however, we can find an explicit distribution of the RV $\lambda(x \mid g(\theta))$ or a certain function of this RV, and then the subsequent choice of the constant c_γ is, certainly, quite justified.

B. Likelihood ratio principle as a method of testing hypotheses.

One of important problems of statistics is that of testing the validity of a certain hypothesis H regarding the true value of a parameter $\theta = (\theta_1, \ldots, \theta_s)$. We now list some examples of such hypotheses:

(1) $H : \theta = \theta_0$,

(2) $H : \theta_1 = \ldots = \theta_s$,

(3) $H : c_1\theta_1 + \ldots + c_s\theta_s = 0$,

and

(4) $H : f(\theta) = f_0$, where $f(\theta)$ is a certain function of θ, f_0 being its fixed value.

In the general case, the hypothesis H is of the form

$$H : \theta \in \Theta_0.$$

Put

$$\lambda_H(x) = \frac{\max_{\theta \in \Theta_0} L(x \mid \theta)}{\max_{\theta \in \Theta} L(x \mid \theta)}.$$

Under quite general conditions (see the corollary in § 5), for "large" n and $x = (x_1, \ldots, x_n)$,

$$\lambda_H(x) \geqslant c_\gamma$$

with probability $1 - \gamma$, where c_γ is determined from the asymptotic equality

$$P\{\chi^2(s-k) < -2\ln c_\gamma\} = 1 - \gamma$$

(s and k are the dimensions of Θ and Θ_0, respectively). The confidence coefficient $1 - \gamma$ is usually given the values 0.9, 0.95, 0.99, or 0.995. E.g., let $1 - \gamma = 0.99$. If the hypothesis H is valid, then $\lambda_H(x) \geqslant c_\gamma$ in 99% of cases. Thus, if we obtained that $\lambda_H(x) < c_\gamma$, which may occur only once out of one hundred times (if the hypothesis H is valid), then we reject the hypothesis H as improbable. If, however, we obtained $\lambda_H(x) \geqslant c_\gamma$, then we can only say that the hypothesis H does not contradict the result $x = (x_1, \ldots, x_n)$ of trials (in practice, the hypothesis H is accepted in this case).

Thus, the condition $\lambda_H(x) \geqslant c_\gamma$ is, with probability $1 - \gamma = 0.9$, only a necessary (but not sufficient) condition for the validity of H. It is only in this non-categorical formulation that we can make use of the so-called tests of goodness of fit (criteria for testing hypotheses). Thus, statistical methods provide for good grounds to reject a hypothesis, and unconvincing reasons to accept it.

Chapter 14

Linear Model Multidimensional Analysis

§ 1. Introduction

It can be noticed that (numerical or vector-valued) RV's frequently used in applications are linearly expressible in terms of RV's with known distributions. If an RV ξ has a distribution with density $p(x)$, then the density of distribution of the RV

$$\eta = a\xi + \mu$$

equals

$$|\det a|^{-1} p[a^{-1}(x - \mu)],$$

where a is a non-singular matrix, and ξ, μ, η vectors of the same dimension.

In this chapter, we shall take up a statistical study related to obtaining information about $\theta = (\mu, a)$ from an observation x, where x is the realization of an RV with distribution density

$$L(x \mid \theta) = |\det a|^{-1} p[a^{-1}(x - \mu)],$$
$$\theta = (\mu, a).$$

In abbreviated form, we will write $\det a = |a|$ and $|\det a| = \|a\|$ (though the notation $\|a\|$ coincides with the accepted notation for the norm of a matrix a).

Example. Let $x_1, \ldots, x_n$ be independent observations having an r-dimensional normal distribution $N_r(m, \sigma\sigma')$. On the basis of $x_1, \ldots, x_n$, we have to make conclusions regarding m and σ. We show that this case is reduced to the linear model just suggested, for which we put

$$x = (x_1, \ldots, x_n).$$

Note that x is actually an rn-dimensional vector with numerical components. The vector x can be represented in the form

$$x = ay + \mu,$$

where $\mu = (\underbrace{m, \ldots, m}_{n \text{ times}})'$ is a column vector, and

$$a = \begin{bmatrix} \sigma & 0 & \vdots & 0 \\ 0 & \sigma & \vdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \vdots & \sigma \end{bmatrix}, \quad \sigma = \{\sigma_{ij}\},$$

whereas the random vector $y = (y_1, \ldots, y_n)'$ has distribution density

$$p(y) = \prod_{k=1}^{n} \frac{1}{(2\pi)^{r/2}} \exp\left\{-\frac{1}{2}(y_k, y_k)\right\}.$$

Then the random vector x has density

$$L(x|\theta) = \|a\|^{-1} p[a^{-1}(x - \mu)], \quad \theta = (\mu, a).$$

Example. If, however, each of x_k has the form $x_k = \sigma y_k + m$, and y_k has density $f(y_k)$ in the above example, then, putting

$$p(y) = f(y_1) \ldots f(y_n),$$

we derive that the density of distribution of the RV x_k equals $\|\sigma\|^{-1} f(\sigma^{-1}(x_k - m))$; therefore, the random vector x has distribution density

$$L(x \mid \theta) = \|a\|^{-1} p[a^{-1}(x - \mu)], \quad \theta = (\mu, a).$$

§ 2. Assumptions *re* Space of Parameters $\theta = (\mu, a)$

Besides the assumption that the desired random vector x has the form

$$x = ay + \mu,$$

where a and μ are a matrix and vector of corresponding dimensions, and y a random vector with known distribution, it is additionally re-

quired in applications that certain restrictions should be placed on the matrix a and vector μ. We make the following assumptions.

1. The sample space X, i.e., space of values taken by the RV x coincides with the Euclidean space R^n.

2. The distribution of the RV x possesses density of the form

$$L(x \mid \theta) = \| a \|^{-1} p[a^{-1}(x - \mu)], \quad \theta = (\mu, a),$$

where $\mu \in M \subset X = R^n$, and $a \in A$ is a certain subset of the set of all real matrices of order $n \times n$.

3. M is a linear manifold in X.

4. A is a certain non-empty group of matrices, i.e.,

(a) it follows from $a \in A$ that there exists an inverse matrix $a^{-1} \in A$, and

(b) it follows from $a, b \in A$ that $ab \in A$.

In particular, the unit matrix $e \in A$, since, if $a \in A$, we obtain $a^{-1} \in A$, and $e = aa^{-1} \in A$.

5. M and A are related by

$$\mu \in M, \quad a \in A \Rightarrow a\mu \in M.$$

Moreover, it is assumed that if Y is the orthogonal complement of M, i.e., $X = M + Y$, then

$$y \in Y, a \in A \Rightarrow ay \in Y.$$

6. An MLE exists and is unique, i.e., for (*almost for* is sufficient) each $x \in X$, the absolute maximum of the function $L(x \mid \theta)$ with respect to $\theta = (\mu, a)$ is attained at a unique point. Besides, the absolute maximum of the function $L(x \mid \mu, a)$ with respect to $\mu \in M$ is attained at a unique point, and also the absolute maximum of the function $L(x \mid \mu, a)$ with respect to $a \in A$ is attained at a unique point.

In the subsequent sections of this chapter, we shall regard these assumptions as valid.

Remark 1. We have assumed that the RV x is of the form

$$x = ay + \mu,$$

where the RV y has a distribution with density $p(y)$. The statements of this chapter remain valid if the RV y has a discrete distribution specified by the probabilities $p(y)$, in which case

$$L(x \mid \theta) = p[a^{-1}(x - \mu)],$$

with obvious changes of the type

$$x \in X, \mu \in M \Rightarrow x - \mu \in X,$$
$$x \in X, a \in A \Rightarrow ax \in X.$$

Remark 2. It may seem that we have not taken into account the case where the RV x does not assume all the values in R^n. E.g., if, in a practical problem, the RV y takes values only in the positive

octant of the space R^n, then we may assume that the density $p(y)$ equals zero outside this octant, and thereby extend the sample space X to the whole space R^n.

§ 3. MLE Property

We have assumed that the MLE $\hat{\theta} = \hat{\theta}(x) = (\hat{\mu}, \hat{a}) = (\hat{\mu}(x), \hat{a}(x))$ defined by the relation

$$L(x \mid \hat{\theta}) = \max_{\theta} L(x \mid \theta)$$

is unique. The basic properties of the functions $\hat{\mu}(x)$ and $\hat{a}(x)$ are expressed by the following.

Lemma.

$$\hat{\mu}(x - \mu) = \hat{\mu}(x) - \mu,$$
$$\hat{a}(x - \mu) = \hat{a}(x),$$
$$\hat{\mu}(ax) = a\hat{\mu}(x),$$
$$\hat{a}(ax) = a\hat{a}(x)$$

for all $x \in X$, $\mu \in M$, $a \in A$.

Proof of the lemma is left to the reader.

§ 4. Estimating Function

We call a numerical function $t(x \mid \theta)$ of observations x and unknown parameter θ **estimating** (parameter θ) if, for each θ, the distribution of the RV $t(x \mid \theta)$ does not depend on θ. In other words, for any number t,

$$P_\theta\{t(x \mid \theta) < t\}$$

does not depend on θ.

If $t(x \mid \theta)$ is an estimating function (and the corresponding distribution has density), then, for any number γ, $0 < \gamma < 1$, we can select two numbers t_1 and t_2 not dependong on θ, so that

$$P_\theta\{t_1 \leqslant t(x \mid \theta) \leqslant t_2\} = \gamma.$$

Then

$$S(x) = \{\theta : t_1 \leqslant t(x \mid \theta) \leqslant t_2\}$$

is an interval estimator with confidence coefficient γ.

Note that the concept of estimating function is naturally extended to the case of a vector function $t(x \mid \theta)$.

Theorem. A numerical function

$$t(x|\theta) = \frac{L(x|\theta)}{\max_{\omega} L(x|\omega)} \tag{1}$$

is an estimating function. Moreover,

$$t(x|\theta) = t[a^{-1}(x-\mu)], \tag{2}$$

where

$$t(x) = \frac{p(x)}{\|\hat{a}\|^{-1} p[\hat{a}^{-1}(x-\hat{\mu})]}. \tag{3}$$

Proof. We have

$$t(x|\theta) = \frac{\|a\|^{-1} p[a^{-1}(x-\mu)]}{\|\hat{a}(x)\|^{-1} p[\hat{a}^{-1}(x)(x-\hat{\mu}(x))]}.$$

We derive from the properties of the functions $\hat{\mu}(x)$ and $\hat{a}(x)$ (see Lemma in § 3) that if we put $y = a^{-1}(x-\mu)$, or $x = ay+\mu$, then

$$|\hat{a}(x)| = |a|\,|\hat{a}(y)|,$$

$$\hat{a}^{-1}(x)(x-\hat{\mu}(x)) = \hat{a}^{-1}(y)\,a^{-1}(ay+\mu-a\hat{\mu}(y)-\mu) = \hat{a}^{-1}(y)(y-\hat{\mu}(y)),$$

from which (2) and (3) follow. Further, let $e(A)$ e equal to 1 or 0 according as Condition A is valid or no .

Then

$$P_\theta\{t(x|\theta) < t\} \int_X e\{t(x|\theta) < t\} L(x|\theta)\,dx$$

$$= \int_X e\{t[a^{-1}(x-\mu)] < t\} \|a\|^{-1} p[a^{-1}(x-\mu)]\,dx$$

$$= \int_X e\{t(y) < t\} p(y)\,dy$$

after the substitution $y = a^{-1}(x-\mu)$. We see that the right-hand side in the chain of the latter equalities does not depend on $\theta = (\mu, a)$, i.e., $t(x \mid \theta)$ is an estimating function.

§ 5. Interval MLE for $\theta = (\mu, a)$

The estimating function $t(x \mid \theta)$ defined by relation (1) of the previous section is the Neyman-Pearson likelihood ratio, and satisfies the inequality

$$0 \leqslant t(x \mid \theta) \leqslant 1$$

for all x and θ. Put

$$L(x) = \max_\theta L(x \mid \theta) = L(x \mid \hat{\theta}(x)).$$

It is clear that

$$S(x) = \{\theta : t(x \mid \theta) \geqslant t\}$$
$$= \{\theta : L(x \mid \theta) \geqslant t\, L(x)\}$$

(for any number t) is an interval MLE. To underline the dependence of the sets $S(x)$ on t, we will denote them by $S(x \mid t)$. It is obvious that

(1) for each $x \in X$, the set $S(x \mid t)$ does not increase with respect to t, i.e.,

$$S(x \mid t_1) \supset S(x \mid t_2) \text{ when } t_1 \leqslant t_2,$$

(2) for each $x \in X$,

$$S(x \mid t) = \Theta \text{ when } t < 0$$
$$S(x \mid t) = \varnothing \text{ when } t > 1,$$

(3) for each number t, the probability

$$\gamma(t) = P_\theta\{\theta \in S(x)\} = P_\theta\{t(x \mid \theta) \geqslant t\} = 1 - \alpha(t)$$

does not depend on θ,

(4) when $\theta = (0, e)$, we obtain that

$$\alpha(t) = 1 - \gamma(t) = P\{t(x) < t\},$$

where the RV x has distribution density $p(x)$, and the function $t(x)$ is given by formula (3), § 4.

(5) the function $\alpha(t)$ is not decreasing, with

$$\alpha(t) = \begin{cases} 0 \text{ when } t \leqslant 0 \\ 1 \text{ when } t > 1 \end{cases}, \text{ and,}$$

(6) determining the number t from the equation

$$\gamma(t) = 1 - \alpha(t) = \gamma$$

for a number γ, $0 < \gamma < 1$, where $\alpha(t)$ has been determined in (4), we get that $S_\gamma(x) = S(x \mid t)$ is an interval MLE with confidence coefficient γ.

§ 6. Independence of $\hat{\mu}$ and $\hat{a}$

If the elements of a random matrix $\hat{a}(x)$ are placed one after another, then we obtain a random vector. Further, making use of the usual concept of independence of random vectors, we can speak, e.g., of the dependence or independence of the random matrix $\hat{a}(x)$ and random vector $\hat{\mu}(x)$. In this section, we shall indicate a necessary and sufficient condition of independence of the RV's $\hat{a}(x)$ and $\mu(x)$ as statistics.

Denote the orthogonal complement of M by Y, i.e.,

$$X = R^n = M \dotplus Y.$$

Any element $x \in X$ is thereby uniquely representable in the form $x = \mu + y$, where $\mu \in M$, $y \in Y$, and $(\mu, y) = 0$.

Theorem. A random matrix $\hat{a}(x)$ and random vector $\hat{\mu}(x)$ are independent if and only if $p(x)$ (a.e.) has the form

$$p(y + \mu) = p_1(y) \cdot p_2(\mu), \quad y \in Y, \ \mu \in M. \tag{1}$$

Remark 1. If $p(x)$ has form (1), then the functions $p_1(y)$ and $p_2(\mu)$ can be defined to be positive semi-definite and the integrals of them on Y and M, respectively, equal to unity. We shall assume that $p_1(y)$ and $p_2(\mu)$ are chosen just in this way; therefore, we can regard them as the densities of RV's with values in Y and M, respectively.

Remark 2. Let $R^n = U \dotplus V$, and $f(x)$ a function integrable with respect to $x \in R^n$. Note the formula

$$\int_{R^n} f(x)\,dx = \int_U \int_V f(u+v)\,du\,dv,$$

where du and dv are volume elements in U and V, respectively. In particular,

$$\int_{X_1} p(x)\,dx = \int_{Y_1} \int_{M_1} p(y+\mu)\,dy\,d\mu$$

for any measurable sets $Y_1 \subset Y$, $M_1 \subset M$ and $X_1 = Y_1 \dotplus M_1$.

Lemma 1. If the RV's $\hat{a}(x)$ and $\hat{\mu}(x)$ are independent for a certain θ, then $\hat{a}(x)$ and $\hat{\mu}(x)$ are independent for any θ.

Proof. (a) Let the RV's $\hat{a}$ and $\hat{\mu}$ be independent when $\theta = \theta_0 = (0, e)$, e being the unit matrix, and μ, a any elements from M and A, respectively. We show that $\hat{a}$ and $\hat{\mu}$ are independent when $\theta = (\mu, a)$.

We have

$$P_{(0,e)}\{\hat{a} \in A_0, \hat{\mu} \in M_0\} = P_{(0,e)}\{\hat{a} \in A_0\} \cdot P_{(0,e)}\{\hat{\mu} \in M_0\}$$

for any (measurable) subsets $A_0 \subset A$ and $M_0 \subset M$, i.e.,

$$\int_{\substack{\hat{a}(y) \in A_0 \\ \hat{\mu}(y) \in M_0}} p(y)\,dy = \int_{\hat{a}(y) \in A_0} p(y)\,dy \cdot \int_{\mu(y) \in M_0} p(y)\,dy.$$

Substituting $y = a^{-1}(x - \mu)$, and taking into account that the inclusions

$$\hat{a}(y) \in A_0, \qquad \hat{\mu}(y) \in M_0$$

are equivalent to

$$\hat{a}(x) \in A_1, \qquad \hat{\mu}(x) \in M_1,$$

where

$$A_1 = aA_0, \qquad M_1 = aM_0 + \mu,$$

we obtain

$$\int_{\substack{\hat{a}(x) \in A_1 \\ \hat{\mu}(x) \in M_1}} L(x|\mu, a)\,dx = \int_{\hat{a}(x) \in A_1} L(x|\mu, a)\,dx \cdot \int_{\hat{\mu}(x) \in M_1} L(x|\mu, a)\,dx,$$

where

$$L(x \mid \mu, a) = \| a \|^{-1} p[a^{-1}(x - \mu)],$$

or, which is the same thing,

$$P_{(\mu,a)}\{\hat{a} \in A_1, \hat{\mu} \in M_1\} = P_{(\mu,a)}\{\hat{a} \in A_1\} \cdot P_{(\mu,a)}\{\hat{\mu} \in M_1\}$$

for any (measurable) subsets $A_1 \subset A$ and $M_1 \subset M$. Therefore, $\hat{a}$ and $\hat{\mu}$ are independent when $\theta = (\mu, a)$.

(b) Conversely, if a, μ are any elements from A, M, respectively, and $\hat{a}$, $\hat{\mu}$ independent when $\theta = (\mu, a)$, then $\hat{a}$, $\hat{\mu}$ are independent when $\theta = (0, e)$. To verify that, we have to retrace our steps in (a).

Lemma 1 follows from the statements of (a) and (b).

Lemma 2. If x has distribution density $p(x)$, and (1) is valid, then the RV's $\hat{a}(x)$ and $\hat{\mu}(x)$ are independent,

Proof. (a) The equality $\max_{\mu} p_2(\mu) = p_2(\mu_0)$ uniquely determines a point $\mu_0 \in M$. Without loss of generality, we may assume that $\mu_0 = 0$. Then $\hat{\mu}(y)$ does not depend on $y \in Y$; moreover, $\hat{\mu}(y) = 0$, which follows from $aM = M$, $aY = Y$, and the definition of $\hat{\mu}(y)$. Thus, $\hat{\mu}(y + \mu) = \hat{\mu}(y) + \mu = \mu$.

(b) Let $A_0 \subset A$, and $M_0 \subset M$. Since $\hat{a}(y + \mu) = \hat{a}(y)$ and $\hat{\mu}(y + \mu) = \mu$ for any $y \in Y$, $\mu \in M$, while every point x is uniquely representable in the form $x = y + \mu$, $y \in Y$, $\mu \in M$, we obtain, putting

$$Y_0 = \{y \in Y, \hat{a}(y) \in A_0\},$$

that

$$\{x : \hat{a}(x) \in A_0, \hat{\mu}(x) \in M_0\} = \{x : x = y + \mu, \ y \in Y_0, \mu \in M_0\} = Y_0 \dotplus M_0,$$

$$\{x : \hat{a}(x) \in A_0\} = Y_0 \dotplus M, \quad \{x : \hat{\mu}(x) \in M_0\} = Y \dotplus M_0,$$

whence

$$P\{\hat{a} \in A_0,\ \mu \in M_0\} = P\{x \in Y_0 + M_0\}$$

$$= \int\limits_{Y_0 \dot{+} M_0} \cdot\, p(x)\, dx = \int\limits_{Y_0} p_1(y)\, dy \int\limits_{M_0} p_2(\mu)\, d\mu$$

$$= \int\limits_{Y_0 \dot{+} M} p(x)\, dx \times \int\limits_{Y \dot{+} M_0} p(x)\, dx = P(\hat{a} \in A_0) \cdot P(\hat{\mu} \in M_0).$$

Lemma 3. Let an RV x have distribution density $p(x)$, and the RV's $\hat{a}(x)$ and $\hat{\mu}(x)$ be independent. Then $p(x)$ has form (1).

Proof. For any (measurable) sets $Y_1 \subset Y$ and $M_1 \subset M$, we put

$$A_0 = \{\hat{a}(x) : x = y + \mu,\ y \in Y_1\},$$

$$M_0 = \{\hat{\mu}(x) : x = y + \mu,\ \mu \in M_1\}.$$

Then the equality

$$P\{\hat{a}(x) \in A_0,\ \hat{\mu}(x) \in M_0\}$$

$$= P\{\hat{a}(x) \in A_0\} \cdot P\{\hat{\mu}(x) \in M_0\}$$

is equivalent to

$$P\{x \in Y_1 \dot{+} M_1\} = P\{x \in Y_1 \dot{+} M\} \cdot P\{x \in Y \dot{+} M_1\}$$

or

$$\int\limits_{Y_1 M_1} \int p(y+\mu)\, dy\, d\mu = \int\limits_{Y_1 M} \int p(y+\mu)\, dy\, d\mu \cdot \int\limits_{Y M_1} \int p(y+\mu)\, dy\, d\mu,$$

whence, putting

$$p_1(y) = \int\limits_M p(y+\mu)\, d\mu, \quad p_2(\mu) = \int\limits_Y p(y+\mu)\, dy,$$

we get

$$\int\limits_{Y_1 M_1} \int p(y+\mu)\, dy\, d\mu = \int\limits_{Y_1 M_1} \int p_1(y)\, p_2(\mu)\, dy\, d\mu,$$

or

$$\int\limits_{Y_1 M_1} \int \varepsilon(y,\ \mu)\, dy\, d\mu = 0$$

for all (measurable) sets $Y_1 \subset Y$, $M_1 \subset M$, whereas

$$\varepsilon(y,\ \mu) = p(y + \mu) - p_1(y) \cdot p_2(\mu),$$

with the consequence that $\varepsilon(y,\ \mu) = 0$ a.e., which is equivalent to (1).

The theorem now follows from Lemmas 1-3.

§ 7. Estimating μ and a Separately

(a) For any $a \in A$, we denote by $\hat{\mu}_a(x)$ a function of x, defined so that

$$\max_{\mu} L(x \mid \mu,\ a) = L(x \mid \hat{\mu}_a(x),\ a).$$

Since the maximum of the function $L(x \mid \mu,\ a)$ with respect to μ is attained at a unique point $\hat{\mu}_a(x)$, $\hat{\mu}_a(x)$ is a single-valued function of x for each $a \in A$.

Similarly, for any $\mu \in M$, we denote by $\hat{a}_\mu(x)$ the function of x, defined so that

$$\max_{a} L(x \mid \mu,\ a) = L(x \mid \mu,\ \hat{a}_\mu(x)).$$

We shall sometimes make use of the notation

$$\hat{\mu}_a = \hat{\mu}_a(x) = \hat{\mu}(x \mid a),\ \hat{a}_\mu = \hat{a}_\mu(x) = \hat{a}(x \mid \mu).$$

The basic properties of the functions $\hat{\mu}(x \mid a)$ and $\hat{a}(x \mid \mu)$ are expressed by the following statement.

Lemma.

$$\hat{\mu}(x - \mu \mid a) = \hat{\mu}(x \mid a) - \mu,$$

$$\hat{\mu}(bx \mid a) = b\hat{\mu}(x \mid b^{-1}a),$$

$$\hat{a}(x - \lambda \mid \mu) = \hat{a}(x \mid \lambda + \mu),$$

$$\hat{a}(bx \mid \mu) = b\hat{a}(x \mid b^{-1}\mu)$$

for all $x \in X$, $\lambda,\ \mu \in M$, $a,\ b \in A$.

Proof is felt to the reader.

(b) Put

$$L(x \mid a) = \max_{\mu} L(x \mid \mu,\ a),$$

$$L(x \mid \mu) = \max_{a} L(x \mid \mu,\ a),$$

$$L(x) = \max_{\mu,a} L(x \mid \mu,\ a),$$

$$t_1(x|a) = \frac{L(x|a)}{L(x)}, \quad t_2(x|\mu) = \frac{L(x|\mu)}{L(x)},$$

$$t(x|\mu,\ a) = \frac{L(x|\mu,\ a)}{L(x)}.$$

Recall that $t(x \mid \mu, a)$ is an estimating function, and

$$t(x \mid \mu, a) = t[a^{-1}(x - \mu)],$$

where $t(x)$ is given by formula (3) in § 4.

Theorem. The functions $t_1(x \mid a)$ and $t_2(x \mid \mu)$ are estimating functions, and

$$(1) \qquad t_1(x \mid a) = t_1(a^{-1}x) = t_1[a^{-1}(x - \mu)],$$
$$t_2(x \mid \mu) = t_2(x - \mu) = t_2[a^{-1}(x - \mu)],$$

where $t_1(x)$ and $t_2(x)$ are given by the relations

$$(2) \qquad t_1(x) = \frac{p[x - \hat{\mu}_e]}{L(x \mid \hat{\mu}, \hat{a})},$$

$$t_2(x) = \frac{\| \hat{a}_0 \|^{-1} p[\hat{a}_0^{-1}x]}{L(x \mid \hat{\mu}, \hat{a})}.$$

In fact, put $y = a^{-1}(x - \mu)$, or, which is the same thing, $x = ay + \mu$. Using Lemmas of § 3 and (a) of this section, we have

$$L(x) = \| \hat{a}(x) \|^{-1} p[\hat{a}^{-1}(x)(x - \hat{\mu}(x))]$$
$$= \| a \|^{-1} \| \hat{a}(y) \|^{-1} p[\hat{a}^{-1}(y)(y - \hat{\mu}(y))],$$
$$L(x \mid a) = \| a \|^{-1} p[a^{-1}(x - \hat{\mu}(x \mid a))]$$
$$= \| a \|^{-1} p[y - \hat{\mu}(y \mid e)],$$
$$L(x \mid \mu) = \| \hat{a}_\mu(x) \|^{-1} p[\hat{a}_\mu^{-1}(x)(x - \mu)]$$
$$= \| a \|^{-1} \hat{a}_0(y) \|^{-1} p[\hat{a}_0^{-1}(y)\, y],$$

from which formulas (1) and (2) follow. The proof that $t_1(x \mid a)$ and $t_2(x \mid \mu)$ are estimating functions is performed in the same way as for the function $t(x \mid \mu, a)$ (see § 4).

(c) Now, we define an interval MLE for μ and a separately on the analogy with $\theta = (\mu, a)$ (see § 5).

The estimating functions $t_1(x \mid a)$ and $t_2(x \mid \mu)$ are likelihood ratios, and satisfy the inequalities

$$0 \leqslant t_1(x \mid a) \leqslant 1, \quad 0 \leqslant t_2(x \mid \mu) \leqslant 1.$$

It is obvious that

$$A(x) = \{a \in A : t_1(x \mid a) \geqslant t\} = \{a : L(x \mid a) \geqslant tL(x)\}$$

is an interval MLE (for any number t). The same is valid for

$$M(x) = \{\mu \in M : t_2(x \mid \mu) \geqslant t\} = \{\mu : L(x \mid \mu) \geqslant tL(x)\}.$$

To underline the dependence of the functions $A(x)$ and $M(x)$ on t, we will denote them by $A(x \mid t)$ and $M(x \mid t)$, respectively. We see that,

(1) for each $x \in X$, the set $A(x \mid t)$ does not increase with respect to t,

$$(2) \qquad A(x \mid t) = A \text{ when } t \leqslant 0,$$
$$A(x \mid t) = \varnothing \text{ when } t > 1,$$

(3) for each number t, probability

$$\gamma_1(t) = P_{(\mu, a)}\{a \in A(x \mid t)\} = P_{(\mu, a)}\{t_1(x \mid a) \geqslant t\} = 1 - \alpha_1(t)$$

does not depend on μ and a,

(4) when $\mu = 0$ and $a = e$, we obtain that

$$\alpha_1(t) = 1 - \gamma_1(t) = P\{t_1(x) < t\},$$

where the RV x has distribution density $p(x)$ and the function $t_1(x)$ is defined in (2),

(5) the function $\alpha_1(t)$ is non-decreasing, with

$$\alpha_1(t) = \begin{cases} 0 \text{ when } t \leqslant 0 \\ 1 \text{ when } t > 1 \end{cases}, \text{ and,}$$

(6) determining the number t from the equation

$$\gamma_1(t) = 1 - \alpha_1(t) = \gamma$$

for γ, $0 < \alpha < 1$, where $\alpha_1(t)$ is defined in (4), we obtain that $A_\gamma(x) = A(x \mid t)$ is an interval MLE with confidence coefficient γ.

Similar properties are also valid for the sets $M(x \mid t)$.

(d) **Theorem.** The RV's $\hat{\mu}(x)$ and $\hat{\mu}_a(x)$ are independent of the RV's $\hat{a}(x)$ and $\hat{a}_\mu(x)$ if and only if $p(x)$ has the form

$$p(y + \mu) = p_1(y) \cdot p_2(\mu),$$

where $X = Y \dotplus M$, $y \in Y$, $\mu \in M$.

This theorem is proved in the same manner as the corresponding theorem of § 6.

Remark. Show that

$$\hat{\mu}(x) = \hat{\mu}_{\hat{a}(x)}(x), \quad \hat{a}(x) = \hat{a}_0[x - \hat{\mu}(x)].$$

Chapter 15

Statistical Hypothesis Testing

§ 1. Hypothesis and Test

Let $(X, \mathcal{A}, \mathcal{P} = \{P_\theta, \theta \in \Theta\})$ be a statistical structure, and x an observation, i.e., a random element from $(X, \mathcal{A})$ with a distribution $P_\theta \in \mathcal{P}$. As always, it is required to make a statement regarding the true value of the parameter θ on the basis of the observation x. Meanwhile, the expression "θ is the true value of a parameter" only means that "the distribution of the RV x is P_θ".

An assumption of the form $\theta \in \Theta_0$, where Θ_0 is a non-empty subset of the set Θ, is called a **hypothesis**, and often denoted by H_0, or H_0: $\theta \in \Theta_0$. We sometimes do in a simpler (and more rational) manner, viz., call any non-empty subset of the set Θ a hypothesis. However, it is the intention to distinguish between the concepts of "assumption" and "set", which makes us employ different symbols H_0 and Θ_0.

A hypothesis $H_0 : \theta \in \Theta_0$ is said to be **simple** if Θ_0 contains one element, and **composite** if a hypothesis is not simple.

Along with the **null** hypothesis $H_0 : \theta \in \Theta_0$, we consider an **alternative** (or **competing**) hypothesis $H_1 : \theta \in \Theta_1$. Meanwhile, we must have $\Theta_0 \cap \Theta_1 = \varnothing$, in which case H_0 and H_1 are often called simply a hypothesis and alternative.

The outcome of any statistical analysis of two hypotheses H_0 and H_1 should be a rule associating an observation x with one of the two decisions, viz., to accept the hypothesis H_0 or alternative H_1. Such a rule is called a **test**. In this connection, we introduce the following definition.

Definition. A test is a statistic with values in $[0, 1]$. A test of a hypothesis H_0 against an alternative H_1 is a test φ ascribing probabilities $1 - \varphi(x)$ and $\varphi(x)$ to the hypotheses H_0 and H_1, respectively.

Thus, if φ is a test of a hypothesis H_0 against an alternative H_1, then the alternative H_1 is accepted on the basis of an observation x with probability $\varphi(x)$ (null hypothesis H_0 is rejected), and the hypothesis H_0 with probability $1 - \varphi(x)$ is accepted.

If $\varphi(x) = 0$ or 1 for all x, then, putting

$$C = \{x \in X : \varphi(x) = 1\} \tag{1}$$

(it is clear that $\varphi(x)$ is the indicator of the set C), we derive that if an observation x is in the set C, then the hypothesis H_1 is accepted; otherwise, the null hypothesis H_0 is accepted. Such a set C is called a **critical region**.

Remark. In the general case, if

$$\varphi(x)[1 - \varphi(x)] = 0 \quad P_\theta\text{-a.c. } \forall \theta \in \Theta_0 \cup \Theta_1,$$

then set (1) is called a critical region, and the test φ **non-randomized.**

§ 2. Power Function. Test Ordering

The significance of a test φ is usually characterized by a **power function** $\beta_\varphi : \Theta \to [0, 1]$ defined by the formula

$$\beta_\varphi(\theta) = E_\theta \varphi(x), \quad \theta \in \Theta$$

(this is the probability to reject the null hypothesis H_0 when the observation x has a distribution P_θ).

The number

$$\alpha_\varphi = \sup_{\theta \in \Theta_0} \beta_\varphi(\theta)$$

is called the **size**, or (significance) **level** of a test φ, and the restriction of the function β_φ to Θ_1 (i.e., the function β_φ considered only on Θ_1) is called the **power** of the test φ.

Acting in accordance with a test φ, we may either come to a correct decision or make one of the two mistakes, viz., reject the null hypothesis H_0 when it is valid (it is an error of the first kind) or accept the null hypothesis H_0 when it is incorrect (this is an error of the second kind). Since we cannot require that a test should simultaneously minimize the probabilities of errors of the first and second kinds, we have to compromise, viz., to bound the probability of an error of the first kind from above and minimize the maximum probability of an error of the second kind.

Definition 1. A test φ of a hypothesis $H_0 : \theta \subset \Theta_0$ against an alternative $H_1 : \theta \in \Theta_1$ is said to be **uniformly most powerful** (u.m.p.), if, for any test ψ, $\alpha_\psi \leqslant \alpha_\varphi \Rightarrow \beta_\psi(\theta) \leqslant \beta_\varphi(\theta)\ \forall \theta \in \Theta$ (and, therefore, $1 - \beta_\varphi(\theta) \leqslant 1 - \beta_\psi(\theta)\ \forall \theta \in \Theta_1$, i.e., the probability of an error of the second kind for φ uniformly with respect to θ is less than or equal to the probability of an error of the second kind for ψ; otherwise, the power of the test φ is greater than or equal to the power of the test uniformly with respect to θ).

Definition 2. A test φ of a hypothesis H_0 against an alternative H_1 is said to be **maximin** if, for any test ψ, $\alpha_\psi \leqslant \alpha_\varphi \Rightarrow \inf_{\theta \in \Theta_1} \beta_\psi(\theta) \leqslant \inf_{\theta \in \Theta_1} \beta_\varphi(\theta)$

(the right-hand inequality means that the maximum probability of an error of the second kind for the test φ is less than or equal to the maximum probability of an error of the second kind for the test ψ; in this connection, the test φ is said to be **minimax**, and not maximin).

It is obvious that a u.m.p. test is maximin, and, in the case of a simple alternative hypothesis H_1, a maximin test is u.m.p.

§ 3. Fundamental Neyman-Pearson Lemma

Assume that two hypotheses H_0 and H_1 are simple, have the form $H_0 : \theta \in \theta_0$, and $H_1 : \theta \in \Theta_1$. Put

$$p_0(x) = \frac{dP_{\theta_0}}{d\mu}(x), \quad p_1(x) = \frac{dP_{\theta_1}}{d\mu}(x),$$

where μ is a certain measure dominating the distributions P_{θ_0} and P_{θ_1} (e.g., $\mu = P_{\theta_0} + P_{\theta_1}$). Note that, in terms of the LF,

$$p_0(x) = L(x \mid \theta_0), \qquad p_1(x) = L(x \mid \theta_1).$$

We assume, in addition, that $P_{\theta_0} \neq P_{\theta_1}$.

Lemma. 1° (*Existence*). For any number $\alpha \in [0, 1]$, there exist a test φ and constant λ such that

(1) $$E_0\varphi(x) = \alpha, \quad \text{where } E_0 = E_{\theta_0},$$

and

(2) $$\varphi(x) = \begin{cases} 1 \text{ when } p_1(x) > \lambda p_0(x) \\ 0 \text{ when } p_1(x) < \lambda p_0(x) \end{cases}.$$

2° (*Sufficiency*). If a test φ satisfies (1) and (2) for a certain λ, then it is a most powerful test of the level α for testing H_0 against an alternative H_1.

3° (*Necessity*). If φ is a most powerful test of the level α and power < 1, then, for a certain λ, (2) holds μ-a.e.

Remark. Since the hypothesis H_1 is simple, it is not neccessary to employ the term "uniformly" in the definition of the u.m.p. test concept.

Proof of the Lemma can be found, e.g. in the book by E. Lehmann *Testing Statistical Hypotheses* (John Wiley & Sons, Inc., N.Y.; Chapman & Hall, Limited, London, 1959).

Chapter 16

Sequential Analysis

§ 1. Problem Posing

A. Let $H_1, \ldots, H_s$ be s inconsistent hypotheses, only one of them being true. To verify which exactly is true, we may perform trials (observations, or experiments).

Let $x_1, \ldots, x_n$ be the results of n such trials. Each of the quantities can be numerical or vector-valued. The influence of the true hypothesis on the results of the sequential observations $x_1, x_2, \ldots$ is manifested by the fact that if H_i is a true hypothesis, then, for each $n \geqslant 1$, the vector $(x_1, \ldots, x_n)$ has a distribution depending on i. Denote the density of this distribution (or probability in the discrete case) by

$$p_{ni} = p_{ni}(x_1, \ldots, x_n).$$

Thus, in the case of independent trials, p_{ni} may have the form

$$p_{ni} = p_i(x_1) \ldots p_i(x_n),$$

where p_i is the density of the distribution of the observation result in each trial if the hypothesis H_i is true.

B. In such a situation, it is required to determine a rule which indicates by the results of n trials if further trials should be stopped or continued and which hypothesis should be taken in the first case (as a true one). Such a rule is called a **sequential decision test** (**rule**, or **procedure**), and denoted by the letter S.

Formally, a sequential decision test (SDT) can be specified, e.g. by the following two equivalent methods.

(1) $$S = \{(\varphi_n, \delta_n)\}_{n \geqslant 1} = \{(\varphi_n(x_1, \ldots, x_n), \delta_n(x_1, \ldots, x_n)\}_{n \geqslant 1}$$

where, for each n, the function $\varphi_n = \varphi_n(x_1, \ldots, x_n)$ only takes the values 0 or 1, whereas $\delta_n = \delta_n(x_1, \ldots, x_n)$ the values $1, \ldots, s$. The meaning of these functions is in the fact that the trials continue until $\varphi_n = 1$ for the first time, and, after these n trials, the hypothesis indexed by δ_n is declared to be true.

(2) $$S = \{(G_n^1, \ldots, G_n^s; G_n)\}_{n \geqslant 1}.$$

Let the results of n trials be equal to $x_1, \ldots, x_n$.

If $(x_1, \ldots, x_n) \subset G_n^i$, then we accept the hypothesis H_i without further observations. If, however, $(x_1, \ldots, x_n) \in G_n$, then we perform the following trial.

We assume that $G_n^i, G_n \subset X^n = X \times \ldots \times X$, and X is the set of values which the result of a trial may take. In addition, the requirements

$$G_1^1 + \ldots + G_1^s + G_1 = X$$

and

$$G_n^1 + \ldots + G_n^s + G_n = G_{n-1} \times X$$

should certainly be fulfilled.

C. We now discuss the question as to how to estimate the consequences of the choice of one or other SDT. In this connection, we consider the following two aspects.

1. The probability α_{ij} of erroneous acceptance of a hypothesis H_i is given if the hypothesis H_j is true for all i and j. We will speak of an SDT S which selects hypotheses with a given error matrix $A = \{\alpha_{ij}\}$ that it is an SDT of power A.

Each SDT S (of power A) induces $E_i(n \mid S)$, the average number of the trials performed when the hypothesis H_i is true. This characteristic permits us to order (partially) the sequential tests of the same power A.

2. Another indicator permitting us to order (partially) sequential tests is related to the losses caused by the cost of each trial, and the incorrect choice of the true hypothesis. Let c be the cost of one trial, and w_{ij} the losses caused by the acceptance of a hypothesis H_i if a hypothesis H_j is true.

If we have selected a certain SDT S, then it is associated with a certain error martix $A = \{\alpha_{ij}\}$, the average losses (risk) being equal to

$$R_j(S) = cE_j(n|S) + \sum_{i=1}^{s} \alpha_{ij} w_{ij},$$

provided that the hypothesis H_j is true.

If, in addition, *a priori* information in the form of a set of probabilities π_j that the hypotheses H_j are true is given, then the Bayes risk (average losses) in employing the SDT S equals

$$R(S) = R_{\pi}(S) = \sum_{j=1}^{s} \pi_j R_j(S).$$

This feature already permits us to order all sequential tests S fully. We can also attain the same by means of a minimax approach, taking the number $R(S) = \max_j R_j(S)$ as the loss indicator due to the use of an SDT S.

§ 2. Probability Ratio Sequential Test for s Hypotheses

A. Select a matrix $K = \{k_{ij}\}_1^s$ with non-negative elements, and construct the following sequential test.

Let the results of the first n trials be $x_1, \ldots x_n$, and

$$p_{ni} = p_{ni}(x_1, \ldots, x_n), \qquad i = 1, \ldots, s.$$

We will say that a hypothesis H_i is preferable over a hypothesis H_j (we will write this fact as $H_i \succ H_j$, or $i \succ j$) if

$$\frac{p_{ni}}{p_{nj}} \geqslant k_{ij}.$$

In order not to have simultaneously that

$$H_i \succ H_j, \text{ and } H_j \succ H_i,$$

we shall require that

(1) $$k_{ij} > k_{ij}^{-1} \text{ when } i \neq j$$

for the elements of the matrix $K = \{k_{ij}\}$. However, if we cannot say which of two hypotheses H_i and H_j is preferable over the other, then we will say that they are incomparable, which may happen in the case where

$$\frac{p_{ni}}{p_{nj}} < k_{ij}, \text{ and } \frac{p_{nj}}{p_{ni}} < k_{ij},$$

which is equivalent to

$$k_{ji}^{-1} < \frac{p_{ni}}{p_{nj}} < k_{ij}.$$

A sequential test can now be determined by the requirement that the trials should go on until one of the hypotheses becomes preferable over the rest; we shall accept just this hypothesis at the epoch when the trials stop. Such a sequential test is called a **probability ratio sequential test** (PRST) of the level $K = \{k_{ij}\}$.

B. We first take up the investigation of the problem as to how to compute the level K of a PRST if the power $A = \{\alpha_{ij}\}$ of this test is given. Recall that α_{ij} is the probability that a hypothesis H_i will be taken if a hypothesis H_j is true.

Theorem.

(2) $$k_{ij} \leqslant \frac{\alpha_{ii}}{\alpha_{ij}}.$$

Proof is broken into the following items.

(a) Let S be a PRST of a level $K = \{k_{ij}\}$. The test S can be written in the form

$$S = \{(G_n^1, \ldots, G_n^s, G_n)\}_{n \geqslant 1},$$

(see the previous section). Then

(3) $$\alpha_{ij} = \sum_{n \geqslant 1} P\{G_n^i | H_j\}.$$

(b) Show that

(4) $$P\{G_n^i | H_i\} \geqslant k_{ij} P\{G_n^i | H_j\}.$$

In fact,

$$P\{G_n^i | H_i\} - k_{ij} P\{G_n^i | H_j\} = \int 1_{G_n^i} [p_{ni} - k_{ij} p_{nj}]\, dx,$$

where $x = (x_1, \ldots, x_n)$. But if $1_{G_n^i}(x) = 1$, i.e., $x \in G_n^i$, then $H_i \succ H_j$ for all $j \neq i$, which means that

$$\frac{p_{ni}}{p_{nj}} \geqslant k_{ij}, \quad \text{or} \quad p_{ni} - k_{ij} p_{nj} \geqslant 0,$$

and (4) follows.

(c) The theorem is now a consequence of (3) and (4).

It is interesting to see what happens if k_{ij} is replaced by $\frac{\alpha_{ii}}{\alpha_{ij}}$. Thus, consider a PRST S' of a level $K' = \{k'_{ij}\}$, where $k'_{ij} = \frac{\alpha_{ii}}{\alpha_{ij}}$. Requirement (1) means that

$$\frac{\alpha_{ii}}{\alpha_{ij}} > \frac{\alpha_{ji}}{\alpha_{jj}}, \quad \text{or} \quad \alpha_{ii}\alpha_{jj} > \alpha_{ij}\alpha_{ji} \text{ for } i \neq j,$$

which is practically always valid. The power of S' will change. Denote it by $A' = \{\alpha'_{ij}\}$.

By the theorem, we should have

$$k'_{ij} \leqslant \frac{\alpha'_{ii}}{\alpha'_{ij}}, \quad \text{or} \quad \frac{\alpha_{ii}}{\alpha_{ij}} \leqslant \frac{\alpha'_{ii}}{\alpha'_{ij}}. \tag{5}$$

Let, e.g., $s = 2$, i.e., we have only two hypotheses H_1 and H_2. Put

$$A = \begin{bmatrix} 1-\alpha & \alpha \\ \alpha & 1-\beta \end{bmatrix}, \quad A' = \begin{bmatrix} 1-\alpha' & \beta' \\ \alpha' & 1-\beta' \end{bmatrix}.$$

Then (5) is written in the form

$$\frac{1-\alpha}{\beta} \leqslant \frac{1-\alpha'}{\beta'}, \quad \text{and} \quad \frac{1-\beta}{\alpha} \leqslant \frac{1-\beta'}{\alpha'},$$

whence we obtain

$$\alpha' \leqslant \frac{\alpha}{1-\beta}, \quad \beta' \leqslant \frac{\beta}{1-\alpha}, \quad \alpha' + \beta' \leqslant \alpha + \beta.$$

Since α and β do not exceed 0.1 in practical problems, the probabilities of the errors α' and β' may exceed those of α and β by a practically insignificant value; moreover, $\alpha' + \beta' \leqslant \alpha + \beta$, which implies in applications that if the power $A = \{\alpha_{ij}\}$ of a test is given, then the level

$$K = \{k_{ij}\}, \quad k_{ij} = \frac{\alpha_{ii}}{\alpha_{ij}}$$

may be taken as that of the corresponding PRST. Note that, for any number of s hypotheses $H_1, \ldots, H_s$, the following inequalities which are the analogues of (6) are corollaries to (5), viz.,

$$\alpha'_{ij} \leqslant \frac{1}{\alpha_{ii}} \alpha_{ij}.$$

C. The next problem to investigate is the question as to whether or not the process of a sequential test ends in a finite number of steps.

If the trials are independent, and p_{ni} are of the form

$$p_{ni} = p_{ni}(x_1, \ldots, x_n) = f_i(x_1) \ldots f_i(x_n),$$

then, putting

$$z_k(i|j) = \ln \frac{f_i(x_k)}{f_j(x_k)}, \quad s_n(i|j) = z_1(i|j) + \ldots + z_n(i|j),$$

$$Z_k = \{z_k(i|j)\}, \quad S_n = s_n(i|j)\},$$

we obtain that the random matrices $Z_1, Z_2, \ldots$ are i.i.d., and the trial process ends when

$$s_n(i \mid j) \geqslant \ln k_{ij}$$

for the first time for all $j = 1, \ldots, s$ and a certain i; meanwhile the hypothesis H_i is selected. In other words, the trials end when the elements of one of the rows of the matrix S_n are not less than the corresponding elements of the matrix

$$L = \{l_{ij}\} = \{\ln k_{ij}\}.$$

Note that the matrices Z_k, S_n are skew-symmetric with zero elements on the principal diagonal. We may always assume that the elements of the principal diagonal of the matrix L are zeroes, since we may always put $k_{ij} = 1$, $i = 1, \ldots, s$.

Thus, in the case of two hypotheses, $s = 2$, and the indicated matrices have the form

$$Z_k = \begin{bmatrix} 0 & z_k \\ -z_k & 0 \end{bmatrix}, \quad S_n = \begin{bmatrix} 0 & s_n \\ -s_n & 0 \end{bmatrix}, \quad L = \begin{bmatrix} 0 & b \\ -a & 0 \end{bmatrix}.$$

Therefore, the trials continue until

$$a < z_1 + \ldots + z_n < b;$$

however, if, for the first time,

$$z_1 + \ldots + z_n \geqslant b \quad \text{or} \quad z_1 + \ldots + z_n \leqslant a,$$

then the hypothesis H_1 or H_2, respectively, is taken. Thus, in the case of sequential testing of two hypotheses by the probability ratio

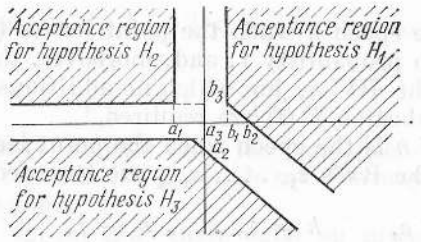

Fig. 16

test, the initial model is equivalent to the model of random walk of a point on an interval (a, b) when the point is shifted through z_k in the k-th step, and the values $z_1, z_2, \ldots$ are i.i.d.

Similarly, in the case of an arbitrary number $s \geqslant 2$ of hypotheses, the initial model can be reduced to the model of walk on a certain domain G of Euclidean space of dimension $s - 1$, the point being shifted through z_k in the k-th step, and the random vectors $z_1, z_2, \ldots$ of dimension $(s - 1)$ being i.i.d.

Thus, in the case $s = 3$, if we put

$$z_k = \left\{\ln \frac{f_1(x_k)}{f_2(x_k)}, \ \ln \frac{f_2(x_k)}{f_3(x_k)}\right\}, \quad L = \begin{bmatrix} 0 & b_1 & b_2 \\ -a_1 & 0 & b_3 \\ -a_2 & -a_3 & 0 \end{bmatrix},$$

then the corresponding walk continues until the point

$$s_n = z_1 + \ldots + z_n$$

enters one of the shaded regions (see Fig. 16).

§ 3. Average Sample Size Finiteness in Using PRST

A. Let a certain point on the straight line vary its position at discrete epochs $n = 1, 2, \ldots$, so that the shift equals z_k at an epoch $n = k$. Then the position of the point at the epoch n equals

$$s_n = z_1 + \ldots + z_n$$

if $s_0 = 0$ at the initial epoch. Assume that the values $z_1, z_2, \ldots$ are i.i.d. with a joint distribution function $F(z)$, and the motion of the point occurs until it leaves the interval (a, b) containing the point 0. Let n be the epoch when the point leaves (a, b) for the first time, i.e.

$$s_n \overline{\in} (a, b); \qquad s_k \in (a, b), \qquad k < n.$$

It is natural to assume that

$$P(z = 0) < 1, \quad \text{i.e.} \quad P(|z| > 0) > 0. \tag{1}$$

We show that the epoch n when the point leaves (a, b) for the first time is finite with probability 1, and, moreover, that there exist all the moments of the RV n, for which no additional assumptions regarding the distribution $F(z)$ are required.

B. Note that if n is the epoch when the point leaves (a, b) for the first time, then the RV's $z_1, \ldots, z_{n-1}$ are bounded, viz.,

$$-c < z_k < c, \qquad k = 1, \ldots, n-1, \qquad c = b - a,$$

from which it follows that if we put

$$z_k' = \begin{cases} -c & \text{for} \quad z_k < -c \\ z_k & \text{for} \quad -c \leqslant z_k \leqslant c, \\ c & \text{for} \quad z_k > c \end{cases} \tag{2}$$

and consider a similar walk process, each time replacing the shifts z_k by z_k', then the epochs n and n' of leaving (a, b) for the first time will coincide for both walk processes, i.e., $n = n'$.

This remark shows that, to investigate the question of the existence of finite moments of the RV n, we can assume from the first that

$$|z_k| \leqslant c = b - a, \qquad k \geqslant 1, \tag{3}$$

with probability 1, and, in particular, that z_k has all the moments.

C. We show that there exists $\varepsilon > 0$ such that $Ee^{\varepsilon n} < \infty$. Hence, it will follow, in particular, that $P(n < +\infty) = 1$, and all the moments of the RV n are finite.

(a) For a certain integer $r \geqslant 1$, we put

$$\zeta_1 = z_1 + \ldots + z_r$$
$$\zeta_2 = z_{r+1} + \ldots + z_{2r}$$
$$\ldots\ldots\ldots\ldots$$
$$\zeta_k = z_{(k-1)r+1} + \ldots + z_{kr}$$
$$\ldots\ldots\ldots\ldots$$

Then the event $n > kr$ is equivalent to the event

$$a < z_1 + \ldots + z_i < b$$

for $i \leqslant kr$; the same is embraced by

$$a < \zeta_1 + \ldots + \zeta_j < b$$

for $j \leqslant k$, whereas the latter event is included in

$$|\zeta_j| < c = b - a$$

for $j \leqslant k$. Thus,

$$P(n > kr) \leqslant P(|\zeta_j| < c, j = 1, \ldots, k) = p^k,$$

where

$$p = P(|\zeta_1| < c).$$

It follows from (1) and (3) that the variance σ^2 of the RV z_1 is positive, and that of the RV ζ_1 equals $r\sigma^2$. We may take r such that $r\sigma^2 > c^2$, and then $p < 1$.

Thus, we have shown that there exist an integer $r \geqslant 1$, and p, $0 \leqslant p < 1$, such that

$$P(n > kr) \leqslant p^k \tag{4}$$

for any integer $k \geqslant 0$.

(b) Let $t > 0$. Then, making use of (4), we obtain

$$\varphi(t) = Ee^{tn} = \sum_{k \geqslant 1} e^{tk} P(n = k) \leqslant \sum_{k \geqslant 0} e^{t(k+1)r} P(kr < n \leqslant (k+1)r)$$
$$\leqslant \sum_{k \geqslant 0} e^{tkr} e^{tr} P(n > kr) \leqslant e^{tr} \sum_{k \geqslant 0} e^{tkr} p^k = e^{tr}(1 - e^{tr}p)^{-1}.$$

Similarly, for $t \leqslant 0$,

$$\varphi(t) \leqslant (1 - e^{tr}p)^{-1},$$

and the corresponding series converge if t is chosen from the condition

$$e^{tr} \cdot p < 1.$$

The convergence interval contains the point $t = 0$, which completes the proof of this item.

D. We now return to the PRST for independent trials when the number s of hypotheses equals 2. We have seen that a model of such

a sequential test is equivalent to the model of walk on the straight line until leaving the interval (a, b) for the first time; meanwhile,

$$z = \ln \frac{f_1(x)}{f_2(x)}, \quad z_k = \ln \frac{f_1(x_k)}{f_2(x_k)},$$
$$a = -\ln k_{21}, \quad b = \ln k_{12},$$

and the RV x has distribution density $f_1(x)$ or $f_2(x)$ according as which of the hypotheses H_1 or H_2 is true. It is natural to assume that the identity $f_1(x) \equiv f_2(x)$ (a.e.) is not true; otherwise, the hypotheses H_1 and H_2 will be indistinguishable, which means that requirement (1) is valid.

The number

$$E_i n = E(n \mid H_i)$$

(equal to the average number of the trials performed) is called the **average sample size**, provided that the hypothesis H_i is true. The statement established in the previous item means that the sample size is finite with probability 1, and, moreover, that all the moments of the random size n are finite.

The proof of the latter statement for the PRST in the case of s hypotheses and independent trials is left to the reader. It only should be assumed that the identity

$$f_i(x) \equiv f_j(x)$$

(a.e.) is valid for no pair of numbers (i, j) such that $i \neq j$ and $i, j = 1, \ldots, s$.

Problems

151. Let an observation x obey a binomial distribution Bi (n, θ) with unknown parameter θ, $0 < \theta < 1$. Assume that the set of possible solutions is $D = R^1$, the loss function is of the form

$$l(d|\theta) = \frac{(d-\theta)^2}{\theta(1-\theta)},$$

and the *a priori* distribution of the parameter θ is uniform on $(0, 1)$. Show that the Bayes decision procedure is unique, and has the form $\delta(x) = x/n$, whereas the Bayes risk equals $1/n$.

152. Assume that the distribution of an observation (x, y) admits density $p(x, y \mid \theta)$, where θ is an unknown parameter. Let

$\alpha(\theta)$ be the density of the *a priori* distribution of the parameter θ,

$\beta(\theta \mid x, y)$ the density of the *a posteriori* distribution of the parameter θ,

$\alpha(\theta \mid x)$ the density of the *a posteriori* distribution with respect to the observation x, and

$\alpha(\theta \mid x, y)$ the density of the *a posteriori* distribution with respect to the observation y if $\alpha(\theta \mid x)$ is taken as the density of the *a priori* distribution.

Show that $\beta(\theta \mid x, y) = \alpha(\theta \mid x, y)$. (Thereby, the *a posteriori* distribution can be calculated in several stages as soon as the results of the observation arrive.)

153. Let $x_1, \ldots, x_n$ be i.i.d. RV's with a distribution from a family $\mathcal{F}$. We will say that the family $\mathcal{F}$ is **consistent** with a family $\mathcal{F}^*$ of distributions on a parameter space if the membership of an *a priori* distribution in $\mathcal{F}^*$ implies that of the *a posteriori* distribution. Show that

(1) the binomial distribution Bi (m, θ) is consistent with the beta distribution,

(2) the negative binomial distribution Bi (m, θ) is consistent with the betadistribution,

(3) the Poisson distribution Po (θ) is consistent with the gamma distribution,

(4) the exponential distribution with parameter θ is consistent with the gamma distribution,

(5) the uniform distribution on $(0, \theta)$ is consistent with the one-sided Pareto distribution, and

(6) the uniform distribution on an interval (θ_1, θ_2) is consistent with the two-sided Pareto distribution.

154. Let $\mathcal{P} = \{P_0, P_1, \ldots, P_m\}$ be a family of distributions on a sample space $(X, \mathcal{A})$. Denote by $p_k(x)$ the density of the distribution P_k with respect to a certain dominating measure μ, e.g., $\mu = P_0 + P_1 + \ldots + P_m$. Assume that $p_0(x) > 0$ for $x \in X$.

Show that the statistic

$$T(x) = \left\{ \frac{p_1(x)}{p_0(x)}, \ldots, \frac{p_m(x)}{p_0(x)} \right\}$$

is sufficient.

155. Let numerical observations $x_1, \ldots, x_n$ be independent, and each have one and the same, but unknown continuous distribution F. Thereby, there is a family $\mathcal{P} = \{P\}$ of distributions of the vector $x = (x_1, \ldots, x_n)$, which correspond to different continuous one-dimensional DF's F. Show that the order statistic

$$T(x) = \{x_{(1)}, \ldots, x_{(n)}\}$$

is sufficient, where $x_{(1)}, \ldots, x_{(n)}$ are the same $x_1, \ldots, x_n$, but placed in non-descending order.

156 (*continued*). Show that an order statistic is complete.

157 (*continued*). Denote an empirical DF by $\hat{F}_x(t)$, i.e.,

$$\hat{F}_x(t) = \frac{1}{n} \sum_{k=1}^{n} e(t - x_k),$$

where

$$e(t)=\begin{cases}0 \text{ for } t\leqslant 0\\ 1 \text{ for } t>0\end{cases}.$$

Show that, for each fixed number t, the estimator $\hat{F}(t)$ is an MVUE of $F(t)$.

158. Consider a statistical structure $(X, \mathcal{A}, \mathcal{P}=\{P_\theta, \theta\in\Theta\})$ in which $X=\{-1, 0, 1, 2, \ldots\}$, $\mathcal{A}$ is the set of all subsets of the set X, $\Theta=(0, 1)$, and

$$P_\theta(-1)=\theta,\ P_\theta(n)=(1-\theta)^2\theta^n,\quad n=0, 1, 2, \ldots.$$

Show that the statistic $T(x)=x$ is boundedly complete, but not complete.

159. Let RV's $x_1, \ldots, x_n$ be independent, and each have a normal distribution $N(\mu, \sigma^2)$, $\theta=(\mu, \sigma^2)$. Show that the statistic

$$T(x)=\left\{\frac{x_1-\bar{x}}{S}, \ldots, \frac{x_n-\bar{x}}{S}\right\}$$

is free, where $n\geqslant 2$, $x=(x_1, \ldots, x_n)$, $\bar{x}=\frac{1}{n}\sum_1^n x_k$, $S=\sum_1^n|x_k-\bar{x}|$.

160 (*continued*). Show that the statistics $\bar{x}$ and S are independent.

Hint: Use the Basu theorem.

161. Let x be the number of successes in a series of n independent trials with constant probability θ of success. Show that

$$t(x)=\sum_{k=0}^{n} c_k\frac{x^{(k)}}{n^{(k)}}$$

is the MVUE for

$$\tau(\theta)=c_0+c_1\theta+\ldots+c_n\theta^n,$$

where

$$x^{(k)}=x(x-1)\ldots(x-k+1).$$

Solution. Due to the completeness of the sufficient statistic $T(x)=x$, it suffices to verify the unbiasedness of the estimator $t(x)$. Put

$$p(z)=E_\theta z^x=\sum_{x=0}^{n}{}^nC_x\theta^x(1-\theta)^{n-x}\cdot z^x=[z\theta+(1-\theta)]^n$$

for $|z|\leqslant 1$. Then

$$\frac{d^k}{dz^k}p(z)\Big|_{z=1}=n^{(k)}\theta^k[z\theta^k+(1-\theta)]^{n-k}|_{z=1}=n^{(k)}\theta^k$$

$$=\sum_{x=0}^{n}{}^nC_x\theta^x(1-\theta)^{n-x}x(x-1)\ldots(x-k+1)z^{x-k}|_{z=1}=E_\theta x^{(k)},$$

i.e., $E_\theta\frac{x^{(k)}}{n^{(k)}}=\theta^k$, $k=0, 1, \ldots, n$, which implies the unbiasedness of the estimator $t(x)$.

162 (*continued*). Show that there exists no UE for $\tau(\theta)=\theta^N$ when $N>n$.

Solution. If $t(x)$ is the required estimator, then

$$E_\theta t(x)=\sum_{k=0}^{n} t(k){}^nC_k\theta^k(1-\theta)^{n-k}=\theta^N,\ 0\leqslant\theta\leqslant 1,$$

but there is a polynomial of degree $n<N$ on the left.

163. Let each of independent RV's $x_1, \ldots, x_n$ have a Poisson distribution Po (θ) with parameter $\theta>0$. Show that the MVUE for $\tau(\theta)=\sum_{k\geqslant 0}a_k\theta^k$ is

$$t(x_1, \ldots, x_n)=\sum_{k=0}^{x} a_k\frac{x^{(k)}}{n^k},$$

where $x=x_1+\ldots+x_n$, $x^{(k)}=x(x-1)\ldots(x-k+1)$.

Solution. Since the statistic $x=x_1+\ldots+x_n$ is complete and sufficient, we may only find a UE for $\tau(\theta)$ as a function of x. Let $t_k(x)$ be a UE for θ^k. Since x is the sum of n RV's which are i.i.d. according to Poisson with parameter θ, the RV x has a Poisson-distribution with parameter $\lambda=n\theta$. We have

$$E_\theta t_k(x)=\sum_{x\geqslant 0} t_k(x)\frac{\lambda^x}{x!}e^{-\lambda}=\theta^k=\frac{1}{n^k}\lambda^k,$$

or

$$\sum_{x\geqslant 0}\frac{t_k(x)}{x!}\lambda^x=\frac{1}{n^k}\lambda^k e^\lambda=\frac{1}{n^k}\sum_{x\geqslant k}\frac{\lambda^x}{(x-k)!},$$

whence, comparing the coefficients of the same powers λ^x, we obtain

$$t_k(x)\begin{cases}=\frac{x!}{n^k(x-k)!}=\frac{x^{(k)}}{n^k} & \text{when } x\geqslant k\\ =0 & \text{when } x<k\end{cases},$$

i.e., $t_k(x)=x^{(k)}/n^k$ (since $x^{(k)}=0$ when $x<k$). It remains to notice that a UE for $\tau(\theta)$ is $t(x)=\sum_{k\geqslant 0}a_k t_k(x)$.

164 (*continued*). Show that, for $\tau(\theta)=1/\theta$, there exists no UE.

Solution. If $t=t(x_1, \ldots, x_n)$ is a UE of $\tau(\theta)$, then

$$E_\theta t=\sum_{x_1\geqslant 0, \ldots, x_n\geqslant 0}\frac{t(x_1, \ldots, x_n)}{x_1!\ldots x_n!}\theta^{x_1+\ldots+x_n}e^{-n\theta}$$

$$=e^{-n\theta}\sum_{k\geqslant 0}c_k\theta^k=1/\theta\quad\forall\theta>0,$$

where

$$c_k = \sum_{\substack{x_1+\ldots+x_n=k \\ x_i \geqslant 0}} \frac{t(x_1, \ldots, x_n)}{x_1! \ldots x_n!}.$$

But the identity

$$\sum_{k \geqslant 0} c_k \theta^k = \frac{1}{\theta} e^{n\theta}$$

is impossible for $\theta > 0$.

165. Given that RV's $x_1, \ldots, x_n$ are i.i.d. according to $N(\mu, \sigma^2)$ with unknown μ and $\sigma > 0$. Put

$$s^2 = \frac{1}{n-1} \sum_1^n (x_k - \bar{x})^2, \quad s \geqslant 0, \quad n > 1, \quad \bar{x} = \frac{1}{n} \sum_1^n x_k.$$

Show that

$$\left(\frac{n-1}{2}\right)^{1/2} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} s$$

is the MVUE for σ.

Hint: See the solution of the previous problem.

166 (*continued*). Show that the MVUE for

$$\tau(\theta) = \sum_{N \geqslant 0} c_N \sigma^N, \quad \theta = (\mu, \sigma),$$

is

$$t(s) = \sum_{N \geqslant 0} \left(\frac{n-1}{2}\right)^{N/2} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{N+n-1}{2}\right)} c_N s^N.$$

Solution. Due to the completeness of the sufficient statistic $T(x) = (\bar{x}, s)$, we shall only prove that

$$E_\theta s^N = \sigma^N \left(\frac{2}{n-1}\right)^{N/2} \frac{\Gamma\left(\frac{N+n-1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}.$$

Put

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_1^n (x_k - \bar{x})^2, \quad \chi^2 \geqslant 0.$$

The RV χ^2 has the chi-square distribution with $m = n - 1$ degrees of freedom. This distribution coincides with that of the RV 2ξ,

where ξ has the gamma distribution $G\left(\frac{m}{2}\right)$, i.e.,

$$P(\chi^2 < t) = \frac{1}{\Gamma\left(\frac{m}{2}\right)} \int_0^{t/2} u^{\frac{m}{2}-1} e^{-u}\, du.$$

Therefore,

$$E\chi^N = E[\chi^2]^{\frac{N}{2}} = \frac{1}{\Gamma\left(\frac{m}{2}\right) 2^{\frac{m}{2}}} \int_0^\infty t^{\frac{N}{2}} t^{\frac{m}{2}-1} e^{-\frac{t}{2}}\, dt$$

$$= \frac{2^{\frac{N+m}{2}} \Gamma\left(\frac{N+m}{2}\right)}{2^{\frac{m}{2}} \Gamma\left(\frac{m}{2}\right)} \int_0^\infty \frac{1}{2^{\frac{N+m}{2}} \Gamma\left(\frac{N+m}{2}\right)} t^{\frac{N+m}{2}-1} e^{-\frac{t}{2}}\, dt.$$

The latter integral equals unity, i.e.,

$$E\chi^N = 2^{\frac{N}{2}} \frac{\Gamma\left(\frac{N+m}{2}\right)}{\Gamma\left(\frac{m}{2}\right)}.$$

It remains to make use of the equality

$$E_\theta s^N = \frac{\sigma^N}{(n-1)^{N/2}} E\chi^N.$$

167*. To determine the reliability of an element whose length of life obeys the exponential distribution $1 - e^{-\theta x}$, $x \geqslant 0$, with unknown parameter $\theta > 0$, a trial with identical N elements is performed up to an epoch T with instantaneous replacement of a faulty element. Show that the MVUE for the reliability $R(t) = e^{-\theta t}$ of the element (t is fixed) is given by the estimator

$$\hat{R}(t) = \left(1 - \frac{t}{NT}\right)^{\nu(T)},$$

where $\nu(T)$ is the number of faulty elements until the trial stops.

Solution. We first verify that the statistic $\nu(T)$ is sufficient and complete.

If $\nu(T) = r$, and $x_1, \ldots, x_r$ are the durations of consecutively placed intervals between neighbouring epochs when the elements get out of order up to the epoch T, then the RV's $x_1, \ldots, x_r$ are independent, and each has the uniform distribution in $[0, T]$ (why?). Therefore, $\nu(T)$ is a sufficient statistic.

The RV $\nu(T)$ has the Poisson distribution with parameter $\lambda = N\theta T$; therefore, the statistic is complete (why?).

It remains to see that $\hat{R}(t)$ is a UE for $R(t)$. Viz.,

$$E_\theta \hat{R}(t) = \sum_{r \geqslant 0} \left(1 - \frac{t}{NT}\right)^r \frac{\lambda^r}{r!} e^{-\lambda} = e^{-\lambda} e^{\lambda - \lambda \frac{t}{NT}}$$

$$= e^{-\lambda \frac{t}{NT}} = e^{-\theta t} = R(t).$$

168. Let RV's $x_1, \ldots x_n$ be independent, and each have the exponential distribution $1 - e^{-\theta x}$, $x \geqslant 0$, with unknown parameter $\theta > 0$. For a fixed $t \geqslant 0$, we put

$$\tau(\theta) = F_\theta(t) = 1 - e^{-\theta t}.$$

Show that the MVUE for $\tau(\theta)$ is

$$F_x^*(t) = 1 - \left(1 - \frac{t}{n\bar{x}}\right)_+^{n-1},$$

when $n > 1$, with $n\bar{x} = x_1 + \ldots + x_n$, $(z)_+ = \max(0, z)$. If, however, $n = 1$, then

$$F_x^*(t) = \begin{cases} 0 \text{ when } x_1 \geqslant t \\ 1 \text{ when } x_1 < t \end{cases}.$$

Solution. The statistic $\bar{x}$ is sufficient and complete. Therefore, it suffices to find a UE for $\tau(\theta)$ as a function of $\bar{x}$. The estimator

$$\delta = \delta(x_1, \ldots, x_n) = \begin{cases} 1 \text{ when } x_1 < t \\ 0 \text{ when } x_1 \geqslant t \end{cases}$$

is a UE for $\tau(\theta)$. Such is also the estimator

$$\delta^*(\bar{x}) = E_\theta\{\delta \mid x_1 + \ldots + x_n = n\bar{x}\}$$

which is, therefore, the required. We have

$$\delta^*(\bar{x}) = P_\theta(x_1 < t \mid x_1 + \ldots + x_n = n\bar{x})$$

$$= P_\theta\left(\frac{x_1}{x_1 + \ldots + x_n} < \frac{t}{n\bar{x}} \,\middle|\, \bar{x}\right).$$

Since the RV's

$$\frac{x_1}{x_1 + \ldots + x_n} \text{ and } x_1 + \ldots + x_n$$

are independent (by the Basu theorem), we get

$$\delta^*(\bar{x}) = P_\theta\left(\frac{x_1}{x_1 + \ldots + x_n} < \frac{t}{n\bar{x}}\right),$$

the right-hand side not depending on θ, and we may assume, in particular, that $\theta = 1$. It remains to note that the RV $x_1/(x_1 + \ldots + x_n)$ has the beta distribution Be $(1, n-1)$. (In fact, if two RV's ξ_1 and ξ_2 are independent and have gamma distributions $G(\lambda_1)$ and $G(\lambda_2)$, then the RV $\xi_1/(\xi_1 + \xi_2)$ has the beta distribution Be (λ_1, λ_2), in which case the RV's $\xi_1 = x_1$ and $\xi_2 = x_2 + \ldots + x_n$, $n > 1$, are independent and have the gamma distributions $G(1)$ and $G(n-1)$, respectively.)

169. To determine the reliability of an element whose length of life obeys the exponential distribution $1 - e^{-\theta x}$, $x \geqslant 0$, with unknown parameter $\theta > 0$, a trial with identical N elements is performed. The trial continues with the replacement of a faulty element until the r-th failure.

Show that the MVUE $R(t) = e^{-\theta t}$ for the reliability of the element (t is fixed) is given by

$$\hat{R}(t) = \left(1 - \frac{t}{s}\right)_+^{r-1},$$

where s is the total length of life of all the elements under trial until it stops.

Solution. Let $y_1, \ldots, y_r$ be the durations of the consecutive intervals between the neighbouring epochs when the elements fail. Then the RV's $y_1, \ldots, y_r$ are independent, and each has the exponential distribution $1 - e^{-N\theta y}$, $y \geqslant 0$ (why?), with

$$s = N(y_1 + \ldots + y_r).$$

Put $x_k = Ny_k$, $k = 1, \ldots, r$. Then the RV's $x_1, \ldots, x_r$ are independent, and each has the distribution $1 - e^{-\theta x}$, $x \geqslant 0$, with

$$s = x_1 + \ldots + x_n.$$

According to Problem 168, the estimator $\hat{R}(t) = 1 - F_s^*(t)$ is the MVUE for $R(t) = 1 - F_\theta(t)$.

170 (*continued*). Prove the same statement, but for the case where the trial is performed without replacing a faulty element ($r \leqslant N$).

Solution. See the solution of the previous problem. In this case, the RV's $y_1, \ldots, y_r$ are independent, and the RV y_k has the exponential distribution

$$1 - \exp\{-(N - k + 1)\theta y\}, \quad y \geqslant 0,$$

with

$$s = Ny_1 + (N-1)y_1 + \ldots + (N - r + 1)y_r.$$

In other respects, the solution is the same if we put

$$x_k = (N - k + 1)y_k, \qquad k = 1, \ldots, r.$$

171. Let $x_1, \ldots, x_n$ be independent RV's having the exponential distribution $1 - e^{-\theta y}$, $y \geqslant 0$, with unknown parameter $\theta > 0$. Show that the MVUE for

$$\tau(\theta) = \sum_{k=0}^{n-1} a_k \frac{\theta^k}{k!}$$

is given by the estimator

$$t(x) = \sum_{k=0}^{n-1} a_k \theta^{n-1} C_k \frac{1}{x^k},$$

where $x = x_1 + \ldots + x_n$.

Solution. Since the statistic $x = x_1 + \ldots + x_n$ is sufficient and complete, then it suffices to show that

$$E_\theta \frac{(n-1)!}{(n-k-1)!} \frac{1}{x^k} = \theta^k, \quad k = 0, 1, \ldots, n-1.$$

Indeed, since the RV θx has the gamma distribution $G(n)$ as an RV equal to the sum of independent RV's θx_k, $k = 1, \ldots, n$, having the gamma distribution $G(1)$, the left-hand side equals

$$\int_0^\infty \frac{(n-1)!}{(n-k-1)!} \cdot \frac{1}{x^k} \theta^n \frac{x^{n-1}}{(n-1)!} e^{-\theta x} dx$$

$$= \theta^k \int_0^\infty \theta^m \frac{x^{m-1}}{(m-1)!} e^{-\theta x} dx = \theta^k, \quad m = n-k \geqslant 1.$$

172 (*continued*). Show that there exists no MVUE for $\tau(\theta) = \theta^N$, when $N \geqslant n$.

Solution. If the required estimator exists, then it depends on $x_1, \ldots, x_n$ through the sufficient statistic $x = x_1 + \ldots + x_n$. Denote this estimator by $t(x)$. The RV θx has the gamma distribution $G(n)$; therefore,

$$E_\theta t(x) = \int_0^\infty t(x) \theta^n \frac{x^{n-1}}{(n-1)!} e^{-\theta x} dx = \theta^N, \quad \theta > 0,$$

or

$$\int_0^\infty f(x) e^{-\theta x} dx = \theta^{N-n}, \quad \theta > 0, \quad N - n \geqslant 0,$$

where $f(x) = t(x) x^{n-1}/(n-1)!$. Differentiate the left-hand and right-hand sides of the latter identity with respect to θ $m = N - n + 1 \geqslant 1$ times (show the correctness of differentiation under the integral sign!). We obtain

$$\int_0^\infty f(x) x^m e^{-\theta x} dx = 0, \quad \theta > 0,$$

whence $f(x) = 0$ a.e. when $x \geqslant 0$, i.e., $t(x) = 0$ a.e. Therefore, $E_\theta t(x) = 0 \neq \theta^N$, $\theta > 0$.

173. Let random observations $x_1, \ldots, x_n$ be independent, and each have a normal distribution $N(\mu, \sigma^2)$ with unknown μ and σ^2.

For a certain number t, put

$$\tau(\theta) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du = F_\theta(t),$$

$\theta = (\mu, \sigma^2)$. Find the MVUE for $\tau(\theta)$.

174. Let random observations $x_1, \ldots, x_m, y_1, \ldots, y_n$ be independent, and, for $t \geqslant 0$,

$$P_\theta(x_i < t) = 1 - e^{-\theta_1 t},$$

$$P_\theta(y_j < t) = 1 - e^{-\theta_2 t},$$

the parameter $\theta = (\theta_1, \theta_2)$ being unknown, $\theta_1 > 0$, $\theta_2 > 0$. Find the MVUE for

$$\tau(\theta) = \frac{\theta_1}{\theta_1 + \theta_2}.$$

175. Consider a Poisson process $\xi(t)$, $t \geqslant 0$, with parameter $\theta > 0$ and $\xi(0) = 0$ (i.e., a homogeneous process $\xi(t)$, $t \geqslant 0$, with independent increments, for which $\xi(0) = 0$, and $\xi(t)$ has the Poisson distribution with parameter θt). Let $x_1, \ldots, x_n$ be observations of the process at given points $t_1 < \ldots < t_n$. Show that $\hat{\theta}(x_1, \ldots, x_n) = x_n/t_n$ is the MLE and MVUE of the parameter θ.

176. Let $x_1, \ldots, x_n$ be observations (values) of a Wiener process with shift coefficient a and diffusion coefficient σ at points $t_1 < \ldots < t_n$. Show that the estimators

$$\hat{a} = \frac{x_n}{t_n} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum_1^n \Delta_k (y_k - \hat{a})^2$$

are the MVUE's for a and σ^2, respectively, where

$$\Delta_k = t_k - t_{k-1}, \quad y_k = \frac{x_k - x_{k-1}}{\Delta_k},$$

$$t_0 = x_0 = 0.$$

177. RV's $x_1, \ldots, x_n$ are i. i. d. according to $N(\mu, \sigma^2)$. Put

$$\bar{x} = \frac{1}{n} \sum_1^n x_i, \quad S = \sum_1^n (x_i - \bar{x})^2,$$

$$s^2 = \frac{1}{n-1} S.$$

Show that, for $n > 3$, $\lambda = 0, 1, 2, 3$ and all $\theta = (\mu, \sigma^2)$,

$$E_\theta \left[\frac{1}{n+\lambda} S - \sigma^2\right]^2 < E_\theta [s^2 - \sigma^2]^2.$$

Solution. Put

$$P(c) = \frac{1}{\sigma^4} E_\theta [cS - \sigma^2]^2 = E_\theta \left[c\frac{S}{\sigma^2} - 1\right]^2,$$

but

$$S/\sigma^2 = \chi^2 (n-1),$$

and, since

$$E\chi^2 (m) = m, \qquad \operatorname{var} \chi^2 (m) = 2m,$$
$$E\,[\chi^2 (n-1)]^2 = n^2 - 1,$$

we obtain

$$P(c) = (n^2-1)\,c^2 - 2(n-1)\,c + 1.$$

The graph of the function $y = P(c)$ is represented in Fig. 17.

Thus, the inequality given in the problem is valid for all $\lambda \in (-1, 3 + 8/(n-3))$.

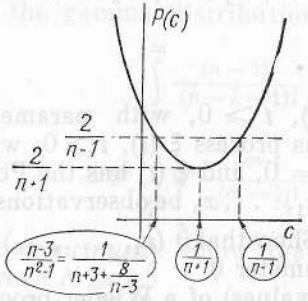

Fig. 17

178 (*continued*). Show that, for any $c \neq 1/(n+1)$ and all θ, the inequality is valid, viz.,

$$E_\theta \left[\frac{1}{n+1} S - \sigma^2\right]^2 < E_\theta\,[cS - \sigma^2]^2.$$

Solution. The minimum of the function $P(c)$ is attained at the point $c = 1/(n+1)$ (see the solution of the previous problem).

179. Let two RV's x_1 and x_2 be independent, and each have a normal distribution $N(0, \sigma^2)$. Show that

$$t(x) = \frac{\sqrt{\pi}}{2} |x_1 + x_2|, \quad x = (x_1, x_2)$$

is a UE of σ.

Solution. The RV $\xi = x_1 + x_2$ has a normal distribution $N(0, \sigma_0^2)$, where $\sigma_0^2 = 2\sigma^2$. It remains to verify that

$$E|\xi| = \sigma_0 \sqrt{\frac{2}{\pi}}.$$

It suffices to verify the latter formula for $\sigma_0 = 1$. We have

$$E|\xi| = \frac{1}{\sqrt{2\pi}} \cdot 2 \int_0^\infty x e^{-\frac{x^2}{2}}\, dx = \sqrt{\frac{2}{\pi}} \int_0^\infty e^{-\frac{x^2}{2}}\, d\frac{x^2}{2} = \sqrt{\frac{2}{\pi}}.$$

180. Assume that two RV's x_1 and x_2 are independent, and each has a normal distribution $N(\mu, \sigma^2)$. Show that

$$t(x) = \frac{\sqrt{\pi}}{2} |x_1 - x_2|, \quad x = (x_1, x_2),$$

is a UE of σ.

Solution. The RV $\xi = x_1 - x_2$ has a normal distribution $N(0, \sigma_0^2)$, where $\sigma_0^2 = 2\sigma^2$. Since $E|\xi| = \sigma_0 \sqrt{\frac{2}{\pi}}$ (see the solution of the previous problem), we obtain

$$E|\xi| = \sqrt{2}\sigma \sqrt{\frac{2}{\pi}} = \frac{2}{\sqrt{\pi}} \sigma.$$

181. Let $x_1, \ldots, x_n$ be i.i.d. RV's with distribution density of the form

$$f(y|\theta) = \begin{cases} k(\theta) \cdot y e^{-\frac{y^2}{\theta^2}} & \text{for } 0 \leqslant y \leqslant \infty \\ 0 & \text{for } y < 0;\ \theta > 0. \end{cases}$$

Show that

$$t(x) = \frac{1}{n} \sum_1^n x_k^2, \quad x = (x_1, \ldots, x_n),$$

is the MVUE for θ^2.

Solution 1. It suffices to show that $t(x)$ is an EE of θ^2. We now find $k(\theta)$. We have

$$1 = \int_0^\infty f(y|\theta)\, dy = k(\theta) \frac{\theta^2}{2} \int_0^\infty e^{-\frac{y^2}{\theta^2}}\, d\frac{y^2}{\theta^2}$$
$$= k(\theta) \frac{\theta^2}{2},$$

whence $k(\theta) = 2\theta^{-2}$. The LF now has the form

$$L(x|\theta) = f(x_1|\theta) \ldots f(x_n|\theta)$$
$$= 2^n \theta^{-2n} (x_1 \ldots x_n) \exp\left\{-\frac{1}{\theta^2} \sum_1^n x_k^2\right\}$$

for $x_1 \geqslant 0, \ldots, x_n \geqslant 0$. Since

$$\frac{\partial}{\partial\theta} \ln L = -\frac{2n}{\theta} + \frac{2}{\theta^3} \sum_1^n x_k^2$$
$$= \frac{2n}{\theta^3} [t(x) - \theta^2] = A(\theta)\,[t(x) - \tau(\theta)],$$

$t(x)$ is an EE of $\tau(\theta) = \theta^2$ $\left(\text{and } \operatorname{var}_\theta t(x) = \frac{\tau'(\theta)}{A(\theta)} = \frac{\theta^4}{n}\right)$.

Solution 2. Since the LF is of the form

$$L(x|\theta) = [k(\theta)]^n (x_1 \ldots x_n) \exp\left\{-\frac{1}{\theta^2} \sum_1^n x_k^2\right\} \cdot 1_{[0,\infty)}(x_1) \ldots 1_{[0,\infty)}(x_n),$$

$t(x)$ is a complete, sufficient statistic. It now suffices to show the unbiasedness of the estimator $t(x)$ with respect to θ^2, i.e., establish that $E_\theta x_1^2 = \theta^2$. Similarly to the previous solution, we find that $k(\theta) = 2\theta^{-2}$. Now,

$$E_\theta x_1^2 = \frac{2}{\theta^2}\int_0^\infty y^2 \cdot y e^{-\frac{y^2}{\theta^2}}\, dy = \theta^2 \int_0^\infty u^2 e^{-u}\, du = \theta^2,$$

where we have put $u = \frac{y^2}{\theta^2}$.

182. Let RV's $x_1, \ldots, x_n$ be independent, and each have a distribution with density

$$f(y|\theta) = \begin{cases} k(\theta)\, y^2 \exp\left\{-\frac{y^3}{\theta^3}\right\} & \text{for } 0 \leqslant y \leqslant \infty \\ 0 & \text{for } y < 0;\ \theta > 0. \end{cases}$$

Show that

$$t(x) = \frac{1}{n}\sum_1^n x_k^3,\ x = (x_1, \ldots, x_n),$$

is the MVUE for θ^3.

Solution is similar to those of the previous problem. Another one may be obtained, proceeding from the following statement, viz., if $\xi_k = x_k^3/\theta^3$, then the RV's $\xi_1, \ldots, \xi_n$ are independent, and each has the exponential distribution $1 - e^{-y}$, $y \geqslant 0$ (for the previous problem, we have to put in this case that $\xi_k = x_k^2/\theta^2$).

183. Let $L(x \mid \theta)$ be the LF of i.i.d. RV's $x_1, \ldots, x_n$; $x = (x_1, \ldots, x_n)$, each being symmetric about the origin, i.e., x_k and $-x_k$ identically distributed. Denote the LF of RV's $y_1, \ldots, y_n$, where $y_k = |x_k|$, $y = (y_1, \ldots, y_n)$, by $M(y \mid \theta)$. Show that

$$M(y \mid \theta) = 2^n L(y \mid \theta)$$

for $y_1 > 0, \ldots, y_n > 0$.

Solution. Denote the distribution and distribution density of the RV x_1 by F_θ and f_θ, respectively, the functions corresponding to the RV y_1 by G_θ and g_θ. Then

$$L(x \mid \theta) = f_\theta(x_1) \ldots f_\theta(x_n),$$
$$M(y \mid \theta) = g_\theta(y_1) \ldots g_\theta(y_n).$$

It suffices to show that $g_\theta(v) = 2f_\theta(v)$, when $v > 0$, which follows from the equalities

$$\begin{aligned} G_\theta(v) &= P(y_1 < v) = P(-v < x_1 < v) = P(x_1 < v) - P(x_1 \leqslant -v) \\ &= P(x_1 < v) - P(-x_1 \geqslant v) \\ &= P(x_1 < v) - [1 - P(-x_1 < v)] \\ &= 2P(x_1 < v) - 1 = 2F_\theta(v) - 1. \end{aligned}$$

184. Let random observations $x_1, \ldots, x_n$ be of the form $x_k = |w_k|^\alpha$, where $\alpha \neq 0$, and the RV's $w_1, \ldots, w_n$ independent and distributed according to $N(0, \theta)$. Find the complete, sufficient statistic and MVUE of the parameter θ.

Solution. The statistic

$$T(x) = \frac{1}{n}\sum_1^n w_k^2 = \frac{1}{n}\sum_1^n x_k^{\frac{2}{\alpha}},\ x = (x_1 \ldots, x_n),$$

is complete, sufficient and the MVUE of the parameter θ.

185. Assume that observations $x_1, \ldots, x_n$ are independent and distributed uniformly on an interval $[0, \theta]$ with unknown **right-hand end** $\theta > 0$. Show that the MVUE $\hat\theta_n$ is consistent, and

$$E_\theta \hat\theta_n = \frac{n}{n+1}\theta,\ \operatorname{var}_\theta \hat\theta_n = \frac{n}{(n+1)^2(n+2)}\theta^2,$$
$$E_\theta(\hat\theta_n - \theta)^2 = \frac{3n+2}{(n+1)^2(n+2)}\theta^2.$$

Solution. The MLE $\hat\theta = \hat\theta_n = \max(x_1, \ldots, x_n)$. Find the distribution of the RV $\hat\theta$, viz.,

$$P_\theta(\hat\theta < u) = P_\theta(x_1 < u, \ldots, x_n < u)$$
$$= \prod_1^n P_\theta(x_k < u) = \left(\frac{u}{\theta}\right)^n,$$

when $0 \leqslant u \leqslant \theta$. Hence,

$$E_\theta\hat\theta = \int_0^\theta u \cdot n\frac{u^{n-1}}{\theta^n}\, du = \frac{n}{n+1}\theta,$$

$$E_\theta\hat\theta^2 = \int_0^\theta u^2 \cdot n\frac{u^{n-1}}{\theta^n}\, du = \frac{n}{n+2}\theta^2,$$

$$P_\theta(|\hat\theta - \theta| \geqslant \varepsilon) \leqslant \frac{E_\theta|\hat\theta-\theta|^2}{\varepsilon^2} = \frac{3n+2}{(n+1)^2(n+2)}\left(\frac{\theta}{\varepsilon}\right)^2,\ \varepsilon > 0.$$

186. Consider a sequence of estimators t_n of a function $\tau(\theta)$ of parameter θ. It is assumed that the estimators t_n and function τ take values in one and the same metric space (E, ρ). Show that if a sequence of the estimators t_n of the function $\tau(\theta)$ is consistent, i.e., $\forall\, \theta$ and $\delta > 0$,

$$P_\theta\{\rho(t_n, \tau(\theta)) \geqslant \delta\} \to 0$$

as $n \to \infty$, then the sequence of the estimators $f(t_n)$ of the function $f(\tau, (\theta))$ is consistent, where f is a continuous mapping of the metric space (E, ρ) into the metric space (M, d).

Solution. $\forall \varepsilon > 0$, $x_0 \in E$ $\exists \delta > 0$ such that $d(f(x), f(x_0)) \geqslant \varepsilon$ implies $\rho(x, x_0) \geqslant \delta$. In particular, if $x_0 = \tau(\theta)$, then

$$P_\theta\{d(f(t_n), f(\tau(\theta))) \geqslant \varepsilon\} \leqslant P_\theta\{\rho(t_n, \tau(\theta)) \geqslant \delta\} \to 0.$$

187. Let $x_1, \ldots, x_n$ be independent RV's distributed according to $N(\mu, \sigma^2)$. Show that $\hat{\theta}_n = (\hat{\mu}, \hat{\sigma})$, where

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_1^n x_k, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_1^n (x_k - \bar{x})^2, \quad \hat{\sigma} \geqslant 0,$$

is the MLE of the parameter $\theta = (\mu, \sigma)$, $\sigma \geqslant 0$. Show also that the estimator $\hat{\theta}_n$ is consistent.

Solution. The first statement follows from $(\hat{\mu}, \hat{\sigma}^2)$ being the MLE for (μ, σ^2), and the bijectivity of the mapping

$$R^1 \times R^1_+ \ni (\mu, \sigma^2) \to (\mu, \sigma) \in R^1 \times R^1_+.$$

The second statement follows from the previous problem.

188. Let RV's $x_1, \ldots, x_n$ be independent, and each have distribution with density of the form $f(y \mid \theta) = e^{-(y-\theta)}$ if $-\infty < \theta \leqslant y < \infty$.

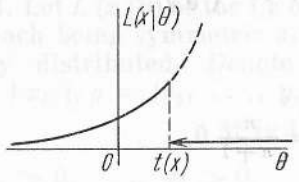

Fig. 18

Show that

$$t(x) = \min(x_1, \ldots, x_n), \quad x = (x_1, \ldots, x_n),$$

is the MLE. Prove the consistency of this estimator.

Solution. Since

$$f(y \mid \theta) = e^{-(y-\theta)} 1_{[0,\infty)}(y), \quad y \in R^1,$$

the LF is represented in the following form (see Fig. 18), viz.,

$$L(x|\theta) = \exp\Big\{-\sum_1^n x_k + n\theta\Big\} \cdot 1_{[\theta,\infty)}(\min x_k)$$

$$= e^{n\theta} 1_{[\theta,\infty)}(t(x)) \cdot \exp\Big\{-\sum_1^n x_k\Big\},$$

from which it follows that $t(x)$ is the MLE. It remains to see that the distribution density of the RV $t(x)$ has the form

$$f_\theta(t) = ne^{-n(t-\theta)} \cdot 1_{[\theta,\infty)}(t),$$

from which we find

$$E_\theta t(x) = \theta + \frac{1}{n}, \quad E_\theta t(x)^2 = \Big(\theta + \frac{1}{n}\Big)^2 + \frac{1}{n^2}, \quad \operatorname{var}_\theta t(x) = \frac{1}{n^2},$$

$$P_\theta\{|t(x) - \theta| \geqslant \varepsilon\} \leqslant \frac{1}{\varepsilon^2} E_\theta[t(x) - \theta]^2 = \frac{2}{(\varepsilon n)^2}, \quad \varepsilon > 0.$$

189. Let RV's $x_1, \ldots, x_n$ be independent, and each have a normal distribution $N(\theta, 2\theta)$. Show that

$$\hat{\theta}_n = \sqrt{1 + \frac{1}{n}\sum_1^n x_k^2} - 1$$

is the MLE. Prove the consistency of this estimator.

Solution. The MLE is found in the usual way among the extremal points of the LF

$$L(x|\theta) = \prod_{k=1}^n \frac{1}{\sqrt{2\pi} \cdot \sqrt{2\theta}} \exp\Big\{-\frac{1}{2 \cdot 2\theta}(x_k - \theta)^2$$

$$= (4\pi)^{-\frac{n}{2}} \theta^{-\frac{n}{2}} \exp\Big\{-\frac{1}{4\theta}\sum_1^n (x_k - \theta)^2\Big\},$$

whence

$$\frac{\partial}{\partial\theta} \ln L = -\frac{n}{2\theta} + \frac{1}{2\theta}\sum_1^n (x_k - \theta) + \frac{1}{4\theta^2}\sum_1^n (x_k - \theta)^2$$

$$= \frac{1}{4\theta^2}\Big[\sum_1^n (x_k - \theta)^2 + 2\theta\sum_1^n (x_k - \theta) - 2\theta n\Big]$$

$$= \frac{1}{4\theta^2}\Big[-n\theta^2 - 2\theta n + \sum_1^n x_k^2\Big]$$

$$= -\frac{n}{4\theta^2}\Big[\theta^2 + 2\theta - \frac{1}{n}\sum_1^n x_k^2\Big].$$

Since the function $L(x \mid \theta)$ vanishes at the ends on the interval $(0, \infty)$ in which θ varies, the MLE is a positive solution of the quadratic equation

$$\theta^2 + 2\theta - \frac{1}{n}\sum_1^n x_k^2 = 0.$$

Let us verify the consistency of the MLE $\hat{\theta}_n$. By the law of large numbers,

$$t_n = \frac{1}{n}\sum_1^n x_k^2 \to E_\theta x_1^2 = \mathrm{var}_\theta x_1 + (E_\theta x_1)^2$$
$$= 2\theta + \theta^2, \quad n \to \infty,$$

in probability. Now, according to Problem 36,

$$\hat{\theta}_n = f(t_n) = \sqrt{1+t_n} - 1 \to f(2\theta+\theta^2)$$
$$= \sqrt{1+2\theta+\theta^2} - 1 = \theta$$

in probability.

190. Let $x_1, \ldots, x_n$ be independent, random, column vectors distributed according to $N_r(\mu, A)$. Show that the MLE $\hat{\theta}_n = (\hat{\mu}, \hat{A})$ of the parameter $\theta = (\mu, A)$ is consistent, where

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_1^n x_k, \quad A = \frac{1}{n}\sum_1^n (x_k - \bar{x})(x_k - \bar{x})'.$$

The solution is based on employing the law of large numbers and representation

$$\hat{A} = \frac{1}{n}\sum_1^n (x_k - \mu)(x_k - \mu)' - (\bar{x} - \mu)(\bar{x} - \mu)'.$$

191. Let $\mathscr{F}$ be the family of distributions F concentrated on $[0, 1]$, and having continuous density functions f, $0 \leqslant f(x) \leqslant c = \text{const}$. Assume that numerical observations $x_1, \ldots, x_n$ are i.i.d. with the same distribution $F \in \mathscr{F}$.

(1) Show that $\hat{F}$ is the MLE for F if and only if $\hat{F} \in \mathscr{F}$ and $\hat{f}(x_i) = c \ \forall\, i$, where $\hat{f}$ is the density of the distribution $\hat{F}$.

(2) Give an example of an inconsistent MLE $\hat{F}$ for F.

(3) Give an example of a consistent MLE $\hat{F}$ for F.

(4) Give an example of an asymptotically normal MLE $\hat{F}$ for F, viz., an MLE $\hat{F}$ such that, for any m functions $g_1, \ldots, g_m$ continuous on $[0, 1]$, the estimator $t(x) = \{t_1(x), \ldots, t_m(x)\}$ is asymptotically normal for $\tau(f) = \{\tau_1(F), \ldots, \tau_m(F)\}$, where $x = (x_1, \ldots, x_n)$,

$$t_i = (g_i, \hat{F}) = \int_0^1 g_i d\hat{F}, \quad \tau_i(F) = (g_i, F) = \int_0^1 g_i\, dF.$$

192. Let random observations $x_1, \ldots, x_n$ be independent and distributed uniformly in $(0, \theta)$ with unknown right-hand end θ. Put

$T(x) = \max(x_1, \ldots, x_n)$. Show that the interval with the ends $T(x)$ and $T(x)\,\varepsilon^{-1/n}$ is an interval estimator of parameter $\theta > 0$ of a level ε, $0 < \varepsilon < 1$.

193. Let $S(x)$ be an interval estimator of a (numerical) parameter θ, and

$$|S(x)| = \int_{S(x)} d\theta$$

the length of the interval $S(x)$. Show that

$$E_{\theta_0}|S(x)| = \int_\Theta P_{\theta_0}\{\theta \in S(x)\}\, d\theta.$$

194. Assume that observations $x_1, \ldots, x_n$ are representable in the form

$$x_k = \theta_0 + \theta_1 t_k + \varepsilon_k, \quad k = 1, \ldots, n,$$

where $t_k = k$, $\theta = (\theta_0, \theta_1)'$ is an unknown parameter, the random observation errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent, and each has a normal distribution $N(0, \sigma^2)$. Denote the LSM-estimator of the parameter θ by $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1)'$.

Show that

$$\mathrm{var}\,\hat{\theta}_1 = \frac{12}{n(n^2-1)}\sigma^2.$$

Solution. 1°. We have

$$x = A\theta + \varepsilon, \quad E\varepsilon = 0, \quad \mathrm{var}\,\varepsilon = \sigma^2 I_n,$$

where

$$x = (x_1, \ldots, x_n)', \quad \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)', \quad \theta = (\theta_0, \theta_1)', \quad A = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}.$$

2°. Recall that $\{\mathrm{cov}\,(\hat{\theta}_i, \hat{\theta}_j)\} = \mathrm{var}\,\hat{\theta} = \sigma^2 (A'A)^{-1}$; therefore, if $\mathrm{var}\,\hat{\theta} = V = \{v_{ij}\}$, then

$$\mathrm{var}\,\hat{\theta}_0 = v_{11}, \quad \mathrm{var}\,\hat{\theta}_1 = v_{22}.$$

3°. We derive

$$W = A'A = \begin{bmatrix} n & \sum t_i \\ \sum t_i & \sum t_i^2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix},$$

$$\sum_1^n t_i = \frac{n(n+1)}{2}, \qquad \sum_1^n t_i^2 = \frac{n(n+1)(2n+1)}{6},$$

$$W^{-1} = \frac{1}{|W|}\begin{bmatrix} w_{22} & -w_{12} \\ -w_{21} & w_{11} \end{bmatrix},$$

whence var $\hat{\theta} = \sigma^2 W^{-1}$; therefore,

$$\operatorname{var} \hat{\theta}_0 = \sigma^2 \frac{1}{|W|} w_{22}, \quad \operatorname{var} \hat{\theta}_1 = \sigma^2 \frac{1}{|W|} w_{11},$$

$$\operatorname{cov} (\hat{\theta}_0, \hat{\theta}_1) = -\sigma^2 \frac{w_{12}}{|W|},$$

in which case we have

$$|W| = w_{11} w_{22} - w_{12}^2 = n \frac{n(n+1)(2n+1)}{6} - \left[\frac{n(n+1)}{2}\right]^2 = \frac{n^2(n^2-1)}{12},$$

whence

$$\operatorname{var} \hat{\theta}_1 = \frac{12}{n(n^2-1)} \sigma^2.$$

195 (*continued*). Show that

$$\operatorname{var} \hat{\theta}_0 = \frac{4}{n}\left[1 + \frac{3}{2(n-1)}\right] \sigma^2.$$

Solution. See the solution of the previous problem. Since

$$\operatorname{var} \hat{\theta}_0 = \frac{w_{22}}{|W|} \sigma^2, \quad w_{22} = \sum_1^n t_i^2 = \frac{n(n+1)(2n+1)}{6},$$

$$|W| = \frac{n^2(n^2-1)}{12},$$

we obtain

$$\operatorname{var} \hat{\theta}_0 = \frac{2(2n+1)}{n(n-1)} \sigma^2 = \frac{4}{n}\left[1 + \frac{3}{2(n-1)}\right] \sigma^2.$$

196 (*continued*). Denote the statistical regression function by

$$f(t) = \hat{\theta}_0 + \hat{\theta}_1 t.$$

Show that

$$Ef(t) = \theta_0 + \theta_1 t, \quad \operatorname{var} f(t) = \frac{\sigma^2}{n} \cdot \frac{\sum_1^n (t_k - t)^2}{\sum_1^n (t_k - \bar{t})^2},$$

where

$$\bar{t} = \frac{1}{n} \sum_1^n t_k.$$

197. Let P_0 be a uniform distribution in the interval (0, 1), P_1 a uniform distribution in the interval (1, 2). Show that the following criteria for testing the distribution P_0 against the competing distribution P_1

$$\varphi_1(x) = \begin{cases} \alpha, & x \in (0, 1) \\ 1, & x \in (1, 2) \end{cases}, \qquad \varphi_2(x) = \begin{cases} 0, & x \in [\alpha, 1) \\ 1, & x \in (0, \alpha) \cup (1, 2) \end{cases},$$

$0 \leqslant \alpha \leqslant 1$, are u.m.p. tests of the level α.

The solution follows from the definition of a u.m.p. test, since, putting

$$E_0 = E_{P_0}, \qquad E_1 = E_{P_1},$$

we have

$$E_0 \varphi_1 = E_0 \varphi_2 = \alpha, \qquad E_1 \varphi_1 = E_1 \varphi_2 = 1.$$

198. Consider a criterion φ for testing a distribution with density p_0 against a competing distribution with density p_1 of the form

$$\varphi(x) = \begin{cases} 1 \text{ when } p_1(x) > k p_0(x) \\ 0 \text{ when } p_1(x) < k p_0(x) \end{cases}.$$

Denote the errors of the first and second kind by α_1 and α_2. Show that

(1) $$\frac{\alpha_2}{1-\alpha_1} \leqslant k \leqslant \frac{1-\alpha_2}{\alpha_1},$$

and

(2) $$\alpha_1 + \alpha_2 \leqslant 1.$$

199 (*continued*). Show that a test φ minimizing $\alpha_1 + \alpha_2$ has the form

$$\varphi(x) = \begin{cases} 1, & p_1(x) > p_0(x) \\ 0, & p_1(x) < p_0(x) \end{cases}.$$

200. Let $(X, A, \mathcal{P} = \{P_\theta, \theta \in \Theta\})$ be a statistical structure, D the space of possible decisions, L the loss function, R the risk function, and Δ the admissible DR class. A class $\Delta_0 \subset \Delta$ is said to be **admissible** if $\forall \delta \in \Delta \exists \delta_0 \in \Delta_0$ such that $R(\delta_0 \mid \theta) \leqslant R(\delta \mid \theta)\ \forall \theta$. Assume that $\Theta = \{\theta_0, \theta_1\}$, $D = \{d_1, d_2\}$, $L(d_i \mid \theta_i) = 0$, $L(d_i \mid \theta_j) > 0$ when $i \neq j$.

Show that the Bayes DR class is admissible.

Index of Definitions

A

Admissible class, 329
Algebra
  minimal, 39
  of sets, 38
  σ-, 39

B

Barrier
  absorbing, 114
  reflecting, 115
Bernstein polynomials, 32

C

Coefficient
  confidence, 280
  diffusion, 182
  shift, 182
Coincidence problem, 16
Comparable elements, 87
Completion, 41
Condition
  consistency, 106
  regularity, 245
Consistent, 86
Contracting chain, 119
Convergence
  almost certain, 47, 159
  in distribution, 160
  in the mean, 159
  in probability, 159
  in the square mean, 159
  weak, 121, 161
Correlation, 23
Covariance, 23
Critical region, 300
Cycle, 148, 208, 224
Cylinder, 107

D

Degrees of freedom, 62
Density
  of a conditional distribution, 85
  of a distribution function, 53
  of probability, 42
Distribution
  a posteriori, 18, 86
  a priori, 18, 86
  arithmetic, 143, 207
  beta, 61
  binomial, 19, 25
  Cauchy, 61
  chi-square, 62
  ergodic, 219
  Erlang, 60
  exponential, 64
  gamma, 60
  hypergeometric, 19
  lattice, 143, 207
  limiting, 120
  logarithmic normal, 64
  multidimensional, 19
  multidimensional hypergeometric, 19
  multidimensional normal, 76
  negative binomial, 27
  non-arithmetic, 143, 207
  non-lattice, 143, 207

Distribution
  normal, 36, 59, 179
  Pearson, 62
  Poisson, 26
  Rayleigh, 64
  regular conditional, 80
  Snedecor-Fisher, 61
  standard, 59
  stationary, 119
  Student, 62,
  t-, 62
  uniform, 64
  Weibull, 64

E

Equation
  normal, 271
  renewal, 220, 141
Estimator
  confidence, 280
  consistent, 261
  efficient, 247
  efficient unbiased, 247
  free interval, 280
  interval, 279
  interval maximum likelihood, 281
  LSM-, 271
  maximum likelihood, 257
  minimum variance unbiased, 240, 243
  multiple, 280
  of a function, 243
  of parameter, 230
  unbiased, 240, 243
Event, 12
  certain, 14
  elementary, 11
  impossible, 14
  $\mathcal{P}$-negligible, 235
  probability of, 12
Events
  equally likely, 43
  equally possible, 43
  equivalent, 86
  independent, 48, 50
  independent classes of, 48
  mutually exclusive, 13
  mutually independent, 14
Evolution, 12
Expectation
  conditional, 76
  mathematical, 20, 51, 181

F

Family
  conjugate, 240
  consistent, 311
  dense, 162
  exponential, 237, 250
  relatively compact, 162
  three-parameter, 64
Fatou lemma, 54
Formula
  Bayes, 86
  Boole, 16
Function
  characteristic, 30, 67
  convex, 72
  correlation, 181, 186
  covariance, 181, 186, 188
  decision, 230
  directly Riemann integrable, 144
  distribution, 52
  elementary, 199
  estimating, 291
  generating, 24
  integrable, 195, 198
  joint distribution, 53
  likelihood, 235
  measurable, 195
  random, 105, 165
  renewal, 141, 204, 220
  risk, 231
  semi-stochastic, 133
  shift, 181
  simple, 195
  spectral, 189
  stationary, 188
  statistical regression, 274
  stochastic, 133
  stochastically continuous random, 165
  structure, 167
  with orthogonal increments, 175

H

Hypothesis, 300
  alternative, 300
  competing, 300
  composite, 300
  null, 300
  preferable, 304
  simple, 300

I

Indicator, 30
Inequality
  Berry-Esseen, 165
  Bhattacharya, 251
  Cauchy-Schwartz, 75
  Chebyshev, 30
  Hölder-Minkowski, 75
  information, 254
  Jensen, 72
  Kolmogorov, 73
  Liapunov, 74
  moment, 74
Integral with respect to measure, 195

L

Law of Large Numbers, 31
  for Markov chains, 125
  Khinchine, 32
Least squares method, 270
Level
  confidence, 180
  of a test, 301
Likelihood ratio, 86
Likelihood ratio principle, 282

M

Markov chain
  aperiodic, 119
  continuous-time, 113
  discrete-time, 112
  homogeneous, 112, 113
  irreducible, 120
  minimal, 118
  standard, 124
Matrix
  covariance, 71
  Fisher information, 254
  one-step transition, 112
  semi-stochastic, 133
  stochastic, 113
Mean value, 20, 181
Mean vector, 70
Measure, 171
  countably additive, 171
  finite, 171
  orthogonal,
  σ-additive, 171
Metric
  Prokhorov, 162
  Prokhorov-Lévy, 161
Minimal property, 139
Moment, 23
  centred, 23
  k-th, 23

O

Operator
  infinitesimal, 127, 132
  stochastic, 121
Ordering, 87

P

Point
  ergodic, 121
  invariant, 122
  of increase, 143
  regeneration, 224
  trivial,
Positive functional, 123
Power of a test, 301
Probability, 39
  conditional, 15
  extension of, 39
  of an elementary event, 12
  of an event, 38
  regular conditional, 80
  subjective, 87
Probability space, 13, 38
  complete, 41
  conditional,
  continuous, 42

Process
  birth-and-death, 118
  Brownian, 216
  Brownian motion, 181
  conservative, 133
  continuous-time, 111
  delayed recurrent renewal, 140
  discrete-time, 111
  equivalent in the wide sense, 106
  equivalent Markov, 127
  Erlang, 221
  Gaussian, 181, 216
  Gaussian-Markov, 217
  homogeneous, 181
  minimal Markov, 118
  non-homogeneous Wiener, 183
  Poisson, 117, 221
  pure birth, 119
  pure death, 119
  random, 105
  recurrent renewal, 140, 220
  regenerative, 149, 210, 224
  renewal, 140
  semi-Markov, 202
  semi-Markov renewal, 202
  standard Wiener, 182
  stationary, 186
  Wiener, 181, 216
  with independent increments, 182
  with orthogonal increments, 175
  with uncorrelated increments, 167
Projection of a measure, 107

R

Random sequence, 105, 111
Random variable, 19, 45
  discrete, 51
  equivalent, 47
  integrable, 51
  real, 19
  summable, 20, 51
  uncorrelated, 24
  vector-valued, 20
Random walk, 307
  free, 114
  with one reflecting barrier, 115
  with two reflecting barriers, 115
Rationally commensurable points, 207
Regeneration epochs, 147
Regression
  exponential, 275
  linear, 274
  polynomial, 275
Renewal intensity, 142
Rule
  decision, 233, 230
  sequential decision, 303
  three-sigma, 280

S

Sequential decision procedure, 303
Semigroup
  contracting, 121
  irreducible, 123
Set
  confidence, 280
  convex, 72
  free, 238
  P-zero, 41
  zero, 41
Size
  average sample, 310
  of a test, 301
Space
  Borel, 81
  equivalent, 81
  of elementary events, 11
  minimal state, 112
  parameter, 229, 234
  phase, 112
  sample, 229
  state, 112
Spectral representation, 189
State
  absorbing, 128, 133
  instantaneous, 133
  recurrent, 203
  stable, 133
Statistic, 235
  complete, 237
  free, 238
  independent, 235
  integrable, 235
  sufficient, 235

Stochastic boundedness, 165
Stochastic uniform continuity, 165
Strategy, 230
  Bayes, 232
  minimax, 232
  optimal, 232
  statistical, 230
Structure
  dominated statistical, 235
  statistical, 234
Symmetric walk, 114

T

Test, 300
  maximin, 301
  minimax, 301
  non-randomized, 300
  probability ratio sequential, 305
  sequential decision, 303
  uniformly most powerful, 301
Theorem
  Basu, 238
  Beppo Levi, 54, 79
  Blackwell, 144, 220
  Bochner-Khinchine, 188
  central limit, 36, 164
Theorem
  completeness, 54
  elementary renewal, 142, 220
  factorization, 236
  Helly, 161
  Herglotz, 188
  key renewal, 144
  Kolmogorov, 183
  Ky Fan, 160
  Lebesgue, 54, 70
  Pólya, 161
  probability extension, 39
  Prokhorov, 162
  Slutsky, 160
  uniqueness, 55, 66
  Weierstrass approximation, 32
Theory
  composite hypothesis testing, 231
  confidence set, 230
  hypothesis testing, 231
  interval estimation, 230
  point estimation, 230
  simple hypothesis testing, 231

V

Variance, 23

To the Reader

Mir Publishers would be grateful for your comments on the content, translation and design of this book. We would also be pleased to receive any other suggestions you may wish to make.

Our address is:
Mir Publishers
2 Pervy Rizhsky Pereulok
I-110, GSP, Moscow, 129820
USSR

Printed in the Union of Soviet Socialist Republics

Also from Mir Publishers

Problems in the Theory of Probability

B. SEVASTYANOV, V. CHISTYAKOV, A. ZUBKOV

The problems presented in the book cover all aspects of the theory of probability usually included in elementary courses. The problems, instructions, solutions and answers are given separately. Principal theoretical propositions are given at the beginning of each section. The book is supplemented by tables necessary for solving the problems.

35